Inspire…Educate…Transform.

# Supervised models

## Time Series Forecasting

Dr. Anand Narasimhamurthy

Thanks to Dr.Sridhar Pappu for the material


**"Prediction is very difficult, especially if it's about the future."**

--Niels Bohr, Nobel laureate in Physics

# Outline

- Basic understanding of time series.
- Overview of main ideas
  - Trend, seasonality, random variations
  - Stationarity, Differencing

- **Auto regressive** (**AR**) models
  - Auto-correlation and partial auto-correlation functions

- **Moving average** (**MA**) models
  - Simple and exponential moving average

- Integrated models : **ARMA, ARIMA**

# What is Time Series data?

- A sequence of data points in successive order, indexed by time.

$$y_t,\ y_{t-1}, y_{t-2}$$

- Eg: Population of the country listed year-wise, Temperature in the city listed by the hour, Number of iPhones sold listed for each quarter

# Forecasting

- Factors needed to forecast the next month's stock price of Tata Motors $(\hat{y}_{t+1})$
  - Current price $(y_t)$
  - Current Sales, Revenue and profit data $(x_1)$
  - Sales trend $(x_2)$
  - Level debt carried by the company $(x_3)$
  - Competition $(x_4)$
  - Import/export rules $(x_5)$
  - Interest rate environment $(x_6)$
  - US/INR exchange rate $(x_7)$
  - Tax rates $(x_8)$
  - Crack down on black money? $(x_9)$
  - Cost of steel? $(x_{10})$
  - Number of smart phones sold? $(x_{11})$

# Forecasting

$$\hat{y}_{t+1} = g(t, x_1, x_2, x_3 \ldots, y_t, y_{t-1}, y_{t-2}, \ldots)$$

$g$ might be some complex linear or nonlinear function.

Time series forecasting attempts to do same forecast just using the past data of $y$, without relying on any other external predictors ($x_i$).

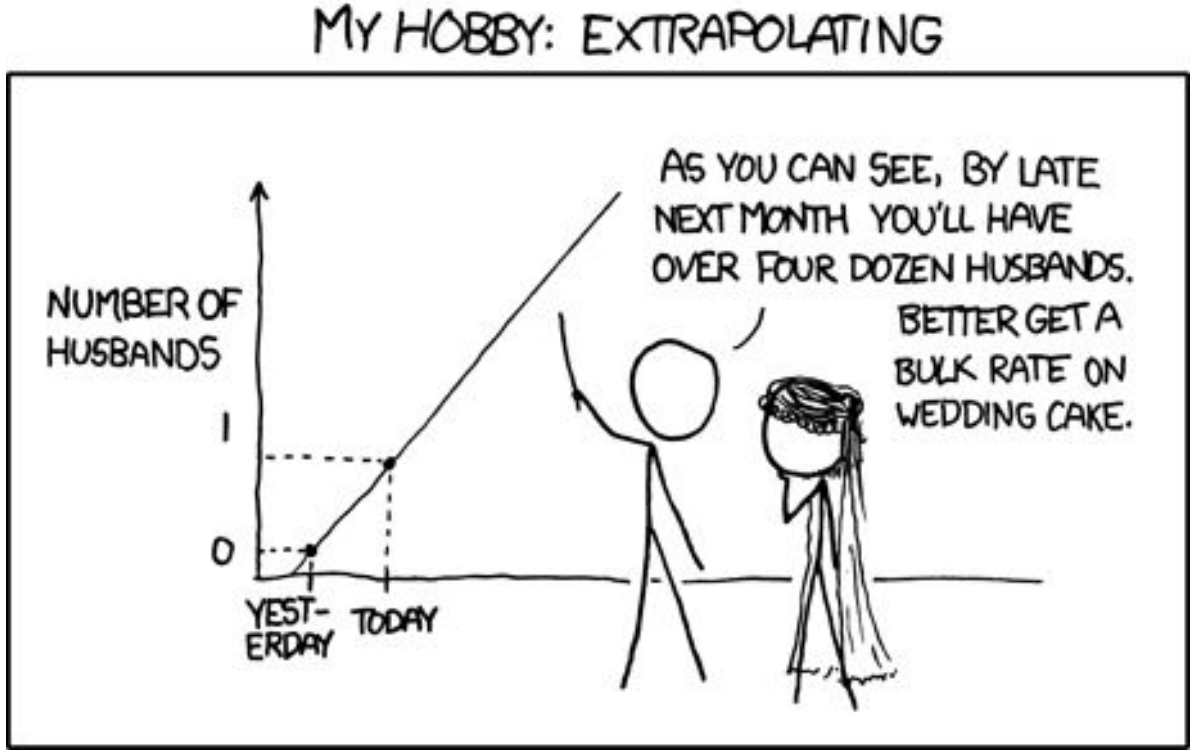
# Typical Time Series

$$\hat{y}_{t+1} = f(t, y_t, y_{t-1}, y_{t-2} \ldots)$$

$f$ can be linear or nonlinear function

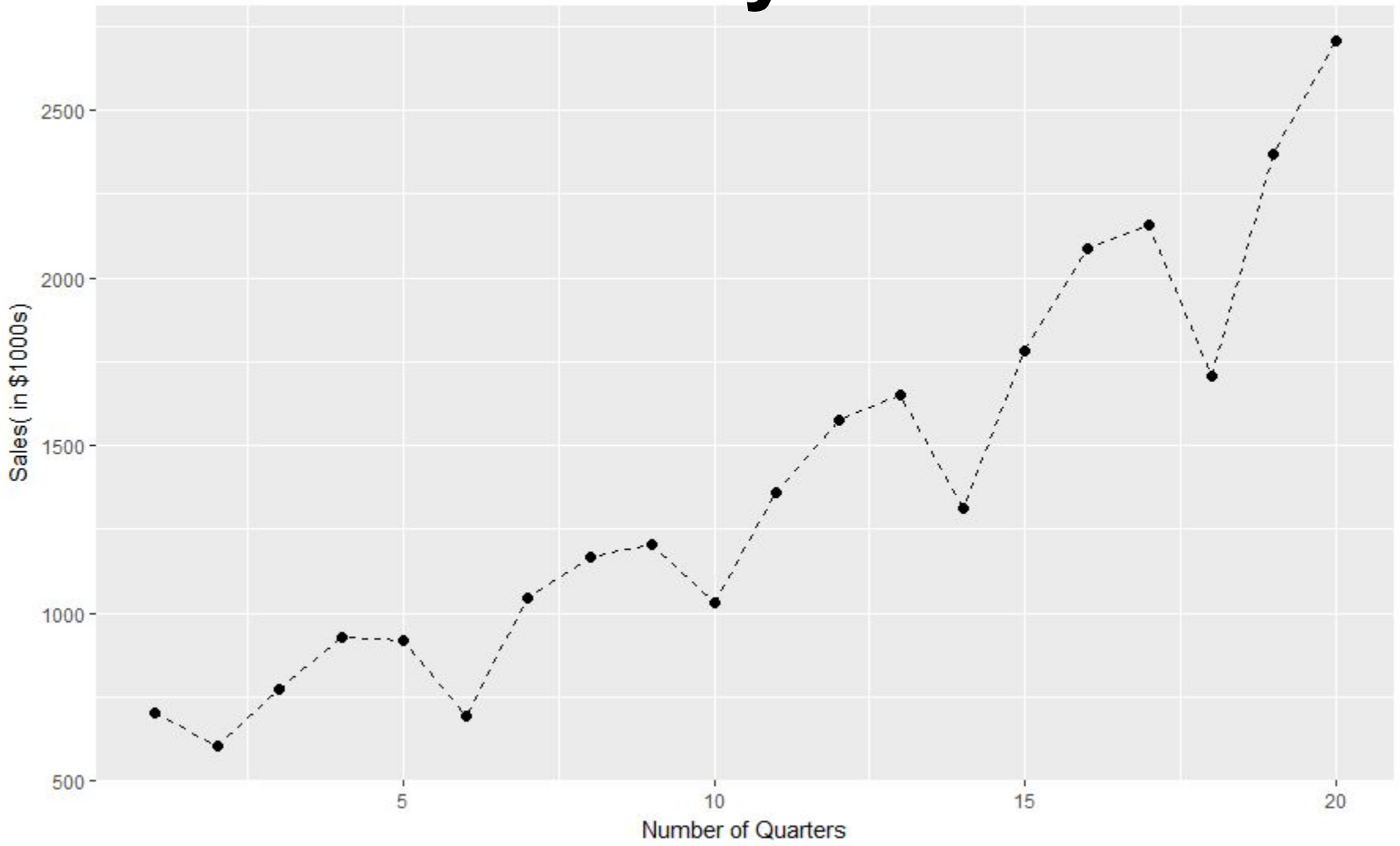
# Why Time Series forecasting?

- Causal independent variables are
  - Unknown to us
  - Not available
  - Might not fit the data well
  - Difficult to forecast




# FORECASTING THROUGH TREND ANALYSIS

# Quarterly sales of a fictitious company : Trend and Seasonality

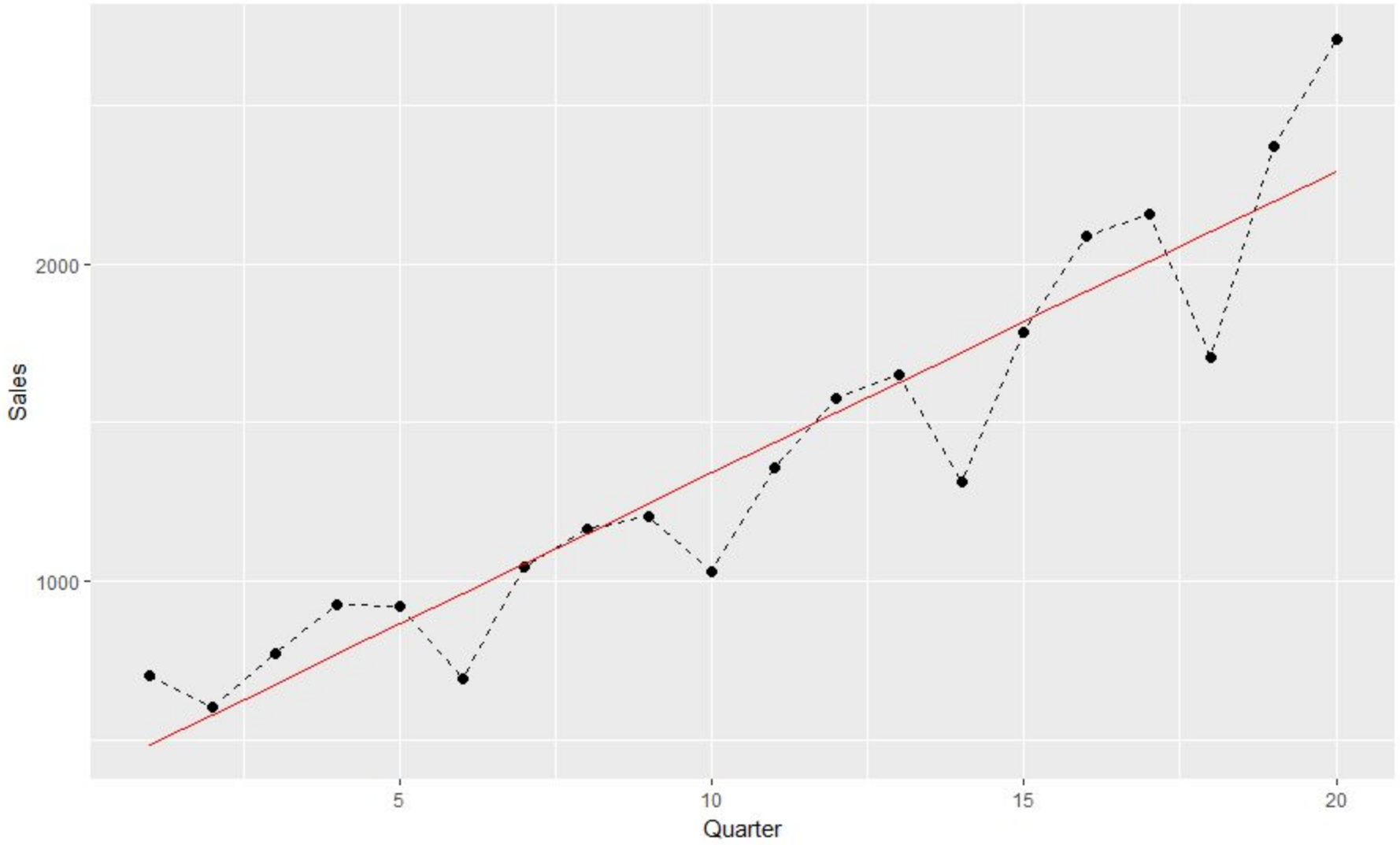


# Regression with time

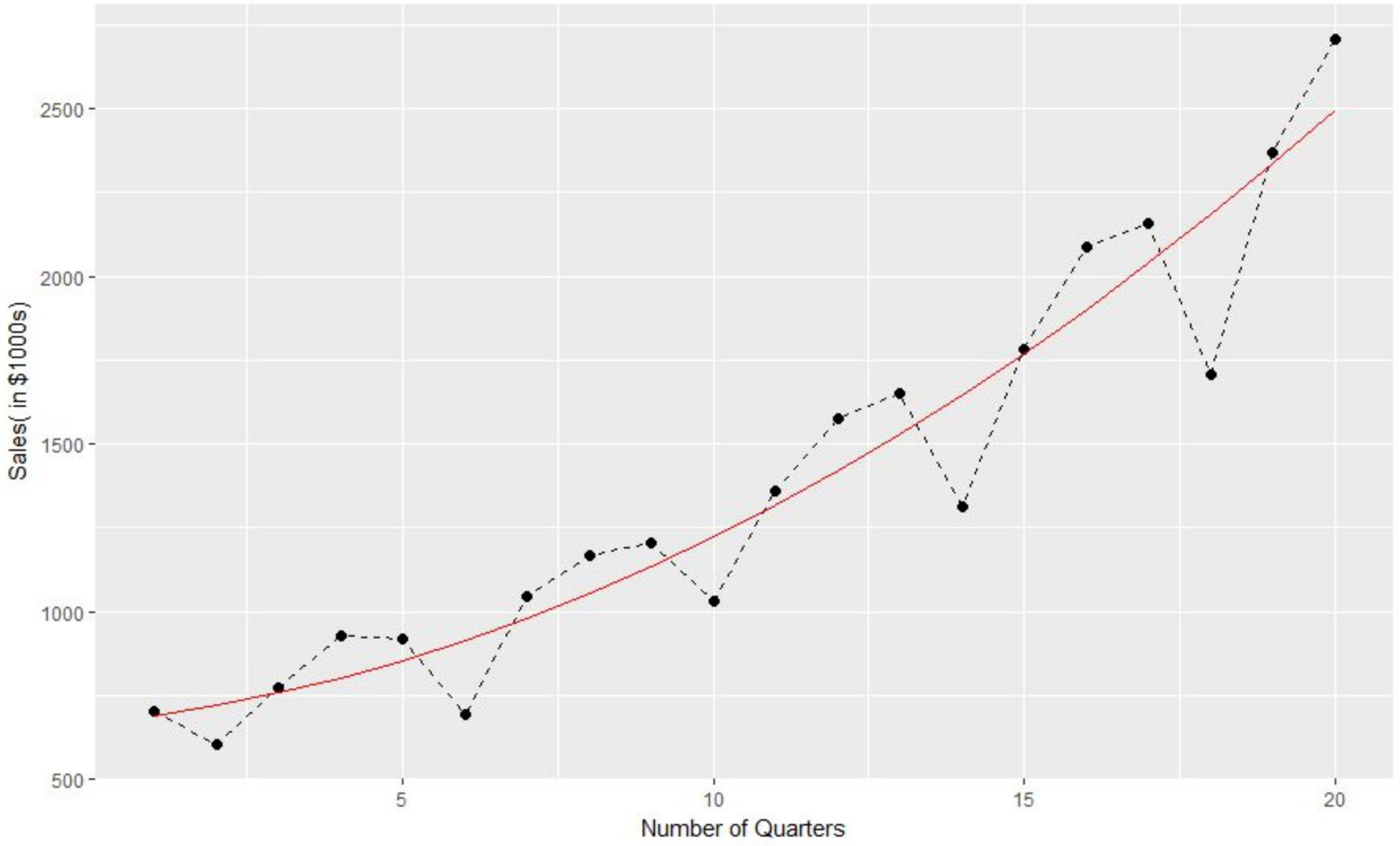
$$\hat{y}_{t+1} = f(t)$$

- Use when trend is the most pronounced component

# Regression Analysis – Linear fit



# Quadratic Trend



# Seasonal Regression Models

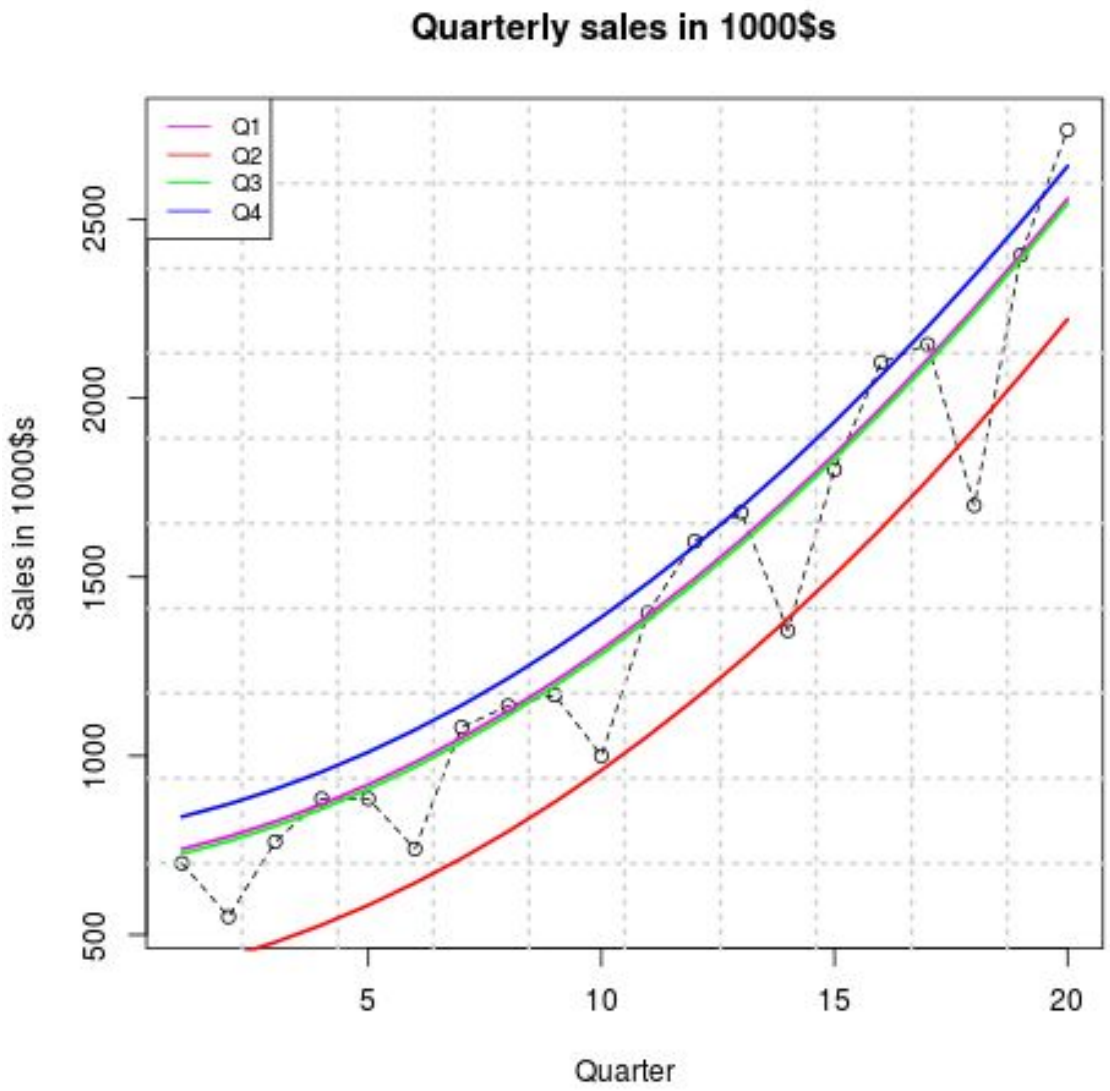
| | Value of | | |
|---|---|---|---|
| Quarter | $X_{3t}$ | $X_{4t}$ | $X_{5t}$ |
| 1 | 1 | 0 | 0 |
| 2 | 0 | 1 | 0 |
| 3 | 0 | 0 | 1 |
| 4 | 0 | 0 | 0 |

$$Y_t = \beta_0 + \beta_1 X_{1_t} + \beta_2 X_{2_t} + \beta_3 X_{3_t} + \beta_4 X_{4_t} + \beta_5 X_{5_t} + \varepsilon_t$$

where, $X_{1t} = t$ and $X_{2t} = t^2$.

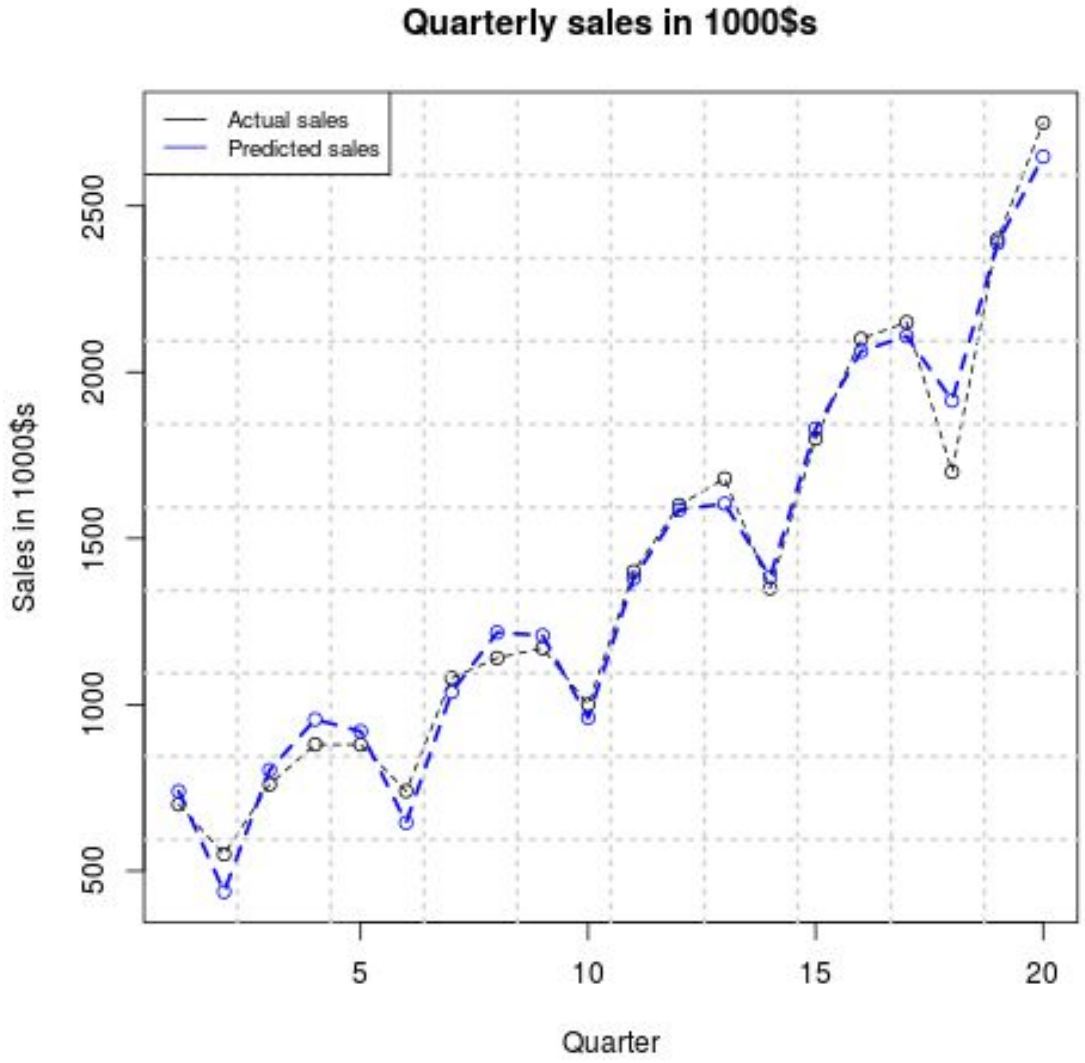
terms including categorical dummy variables coding quarter number.

# Quadratic fit with seasonality




Plotting the fitted curves (+data-points) separately for each quarter, shows fitting a quadratic line for each quarter with a different intercept.

# Seasonal Regression Models



# Goodness of Fit

- MSE (Mean square error)
- MAE (Mean absolute error)
- RMSE (Root mean square error)

- MAPE (Mean absolute percent error)

- NMSE (Normalized mean square error)
- NMAE (Normalized mean absolute error)
- NMAPE (Normalized mean absolute percent error)

# Another Simple Way of Incorporating Seasonality

- Take the trend prediction and actual prediction.
- Depending on **additive** or **multiplicative** model (example in next slides) compute the deviation and map it as seasonality effect for each prediction.
- Take averages of the seasonality value. Use this to make future predictions.

# Case : Quarterly revenues of a company

| Year | Quarter | Time variable (this is created) | Revenues (in $M) |
|---|---|---|---|
| 2008 | I | 1 | 10.2 |
| | II | 2 | 12.4 |
| | III | 3 | 14.8 |
| | IV | 4 | 15 |
| 2009 | I | 5 | 11.2 |
| | II | 6 | 14.3 |
| | III | 7 | 18.4 |
| | IV | 8 | 18 |

```
Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-3.5595 -0.9384  0.4405  1.3265  1.9286

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  10.0393     1.5531   6.464  0.00065 ***
x             0.9440     0.3076   3.069  0.02196 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.993 on 6 degrees of freedom
Multiple R-squared: 0.6109,     Adjusted R-squared: 0.5461
F-statistic: 9.422 on 1 and 6 DF,  p-value: 0.02196
```

What is the Regression equation?

$$y = 10.0393 + 0.9440x$$

# Seasonality: Multiplicative

| Time | Observed values<br>TSI (assuming no impact of cyclicality) | Predicted values (per the regression)<br>T | SI = TSI/T |
|---|---|---|---|
| 1 | 10.2 | 10.983 | 0.929 |
| 2 | 12.4 | 11.927 | 1.040 |
| 3 | 14.8 | 12.871 | 1.150 |
| 4 | 15.0 | 13.815 | 1.086 |
| 5 | 11.2 | 14.759 | 0.759 |
| 6 | 14.3 | 15.703 | 0.911 |
| 7 | 18.4 | 16.647 | 1.105 |
| 8 | 18.0 | 17.591 | 1.023 |

T: Trend; S: Seasonal; I:Irregular

# Quarterly Seasonality

| Time | Average seasonality factor |
|---|---|
| Q1 | 0.844 |
| Q2 | 0.975 |
| Q3 | 1.127 |
| Q4 | 1.054 |

| Time | Observed values<br><br>TSI* (assuming no impact of cyclicality) | Predicted values (per the regression)<br><br>T* | SI* = TSI/T |
|---|---|---|---|
| 1 | 10.2 | 10.983 | 0.929 |
| 2 | 12.4 | 11.927 | 1.040 |
| 3 | 14.8 | 12.871 | 1.150 |
| 4 | 15.0 | 13.815 | 1.086 |
| 5 | 11.2 | 14.759 | 0.759 |
| 6 | 14.3 | 15.703 | 0.911 |
| 7 | 18.4 | 16.647 | 1.105 |
| 8 | 18.0 | 17.591 | 1.023 |

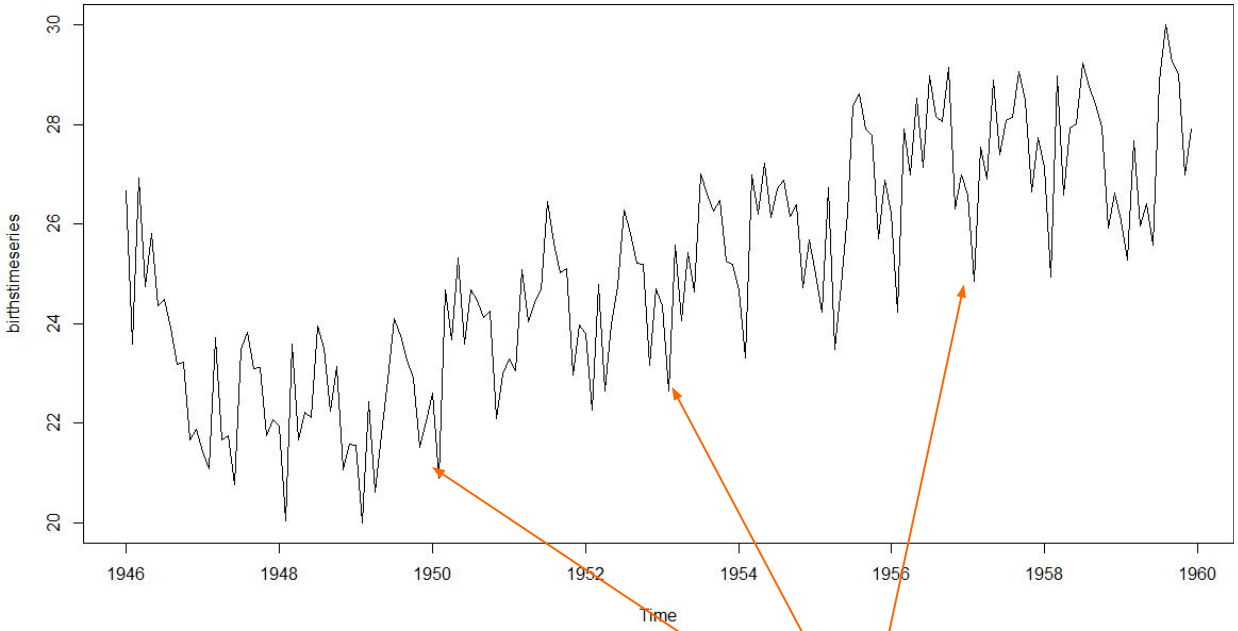
# Computations

- Trend $Y_9$ =10.039 + 0.944(9) = 18.535

- Corrected for seasonality and randomness: 18.535 * 0.844 = 15.643

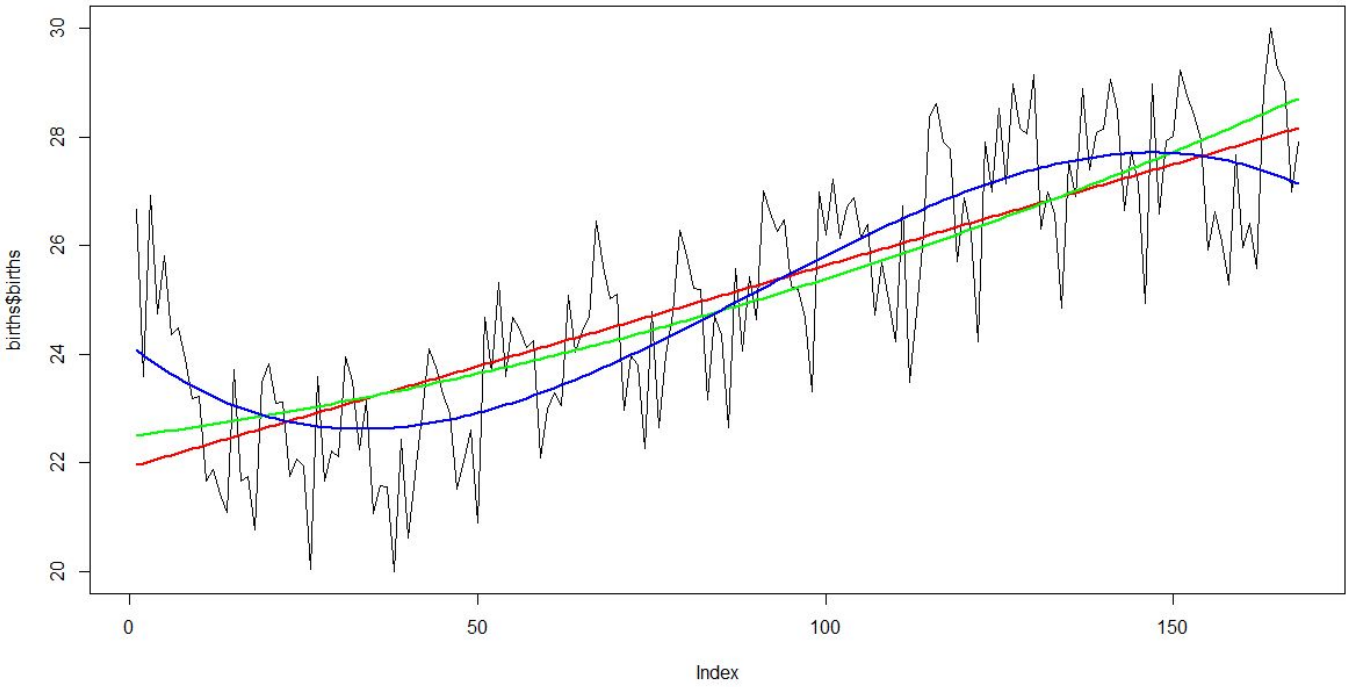
# Seasonal Regression Models

# Births in NY dataset




```
> birthstimeseries <- ts(births,
+                        frequency=12,
+                        start=c(1946,1))
> birthstimeseries
        Jan    Feb    Mar    Apr    May    Jun    Jul    Aug    Sep    Oct    Nov    Dec
1946 26.663 23.598 26.931 24.740 25.806 24.364 24.477 23.901 23.175 23.227 21.672 21.870
1947 21.439 21.089 23.709 21.669 21.752 20.761 23.479 23.824 23.105 23.110 21.759 22.073
1948 21.937 20.035 23.590 21.672 22.222 22.123 23.950 23.504 22.238 23.142 21.059 21.573
1949 21.548 20.000 22.424 20.615 21.761 22.874 24.104 23.748 23.262 22.907 21.519 22.025
1950 22.604 20.894 24.677 23.673 25.320 23.583 24.671 24.454 24.122 24.252 22.084 22.991
1951 23.287 23.049 25.076 24.037 24.430 24.667 26.451 25.618 25.014 25.110 22.964 23.981
1952 23.798 22.270 24.775 22.646 23.988 24.737 26.276 25.816 25.210 25.199 23.162 24.707
1953 24.364 22.644 25.565 24.062 25.431 24.635 27.009 26.606 26.268 26.462 25.246 25.180
1954 24.657 23.304 26.982 26.199 27.210 26.122 26.706 26.878 26.152 26.379 24.712 25.688
1955 24.990 24.239 26.721 23.475 24.767 26.219 28.361 28.599 27.914 27.784 25.693 26.881
1956 26.217 24.218 27.914 26.975 28.527 27.139 28.982 28.169 28.056 29.136 26.291 26.987
1957 26.589 24.848 27.543 26.896 28.878 27.390 28.065 28.141 29.048 28.484 26.634 27.735
1958 27.132 24.924 28.963 26.589 27.931 28.009 29.229 28.759 28.405 27.945 25.912 26.619
1959 26.076 25.286 27.660 25.951 26.398 25.565 28.865 30.000 29.261 29.012 26.992 27.897
>
```

# Seasonal Regression Models - Births

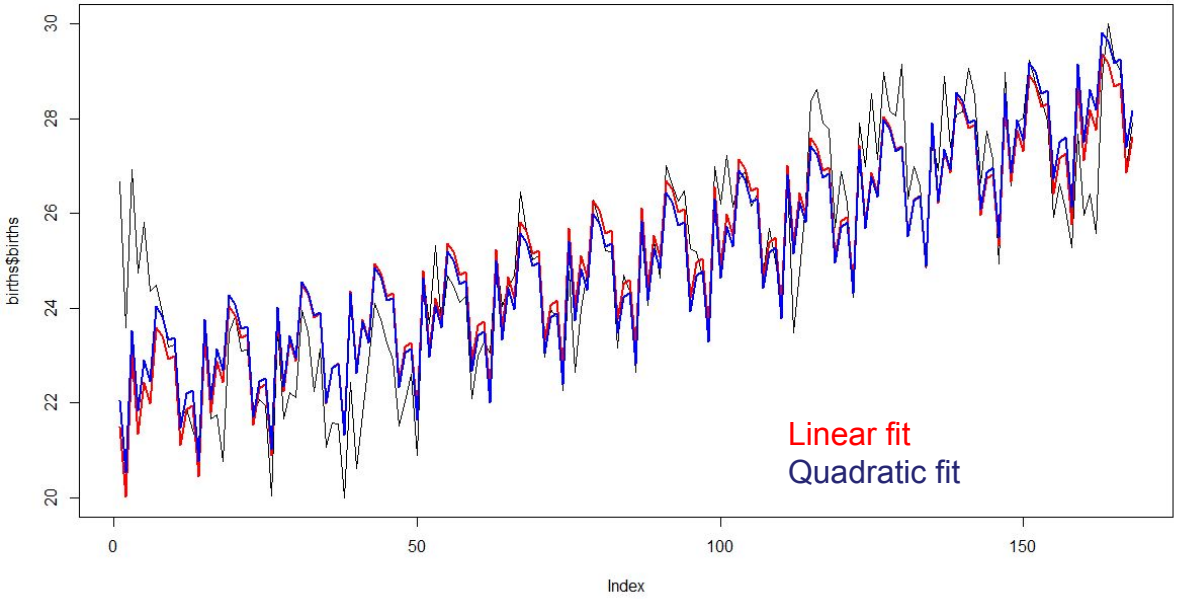


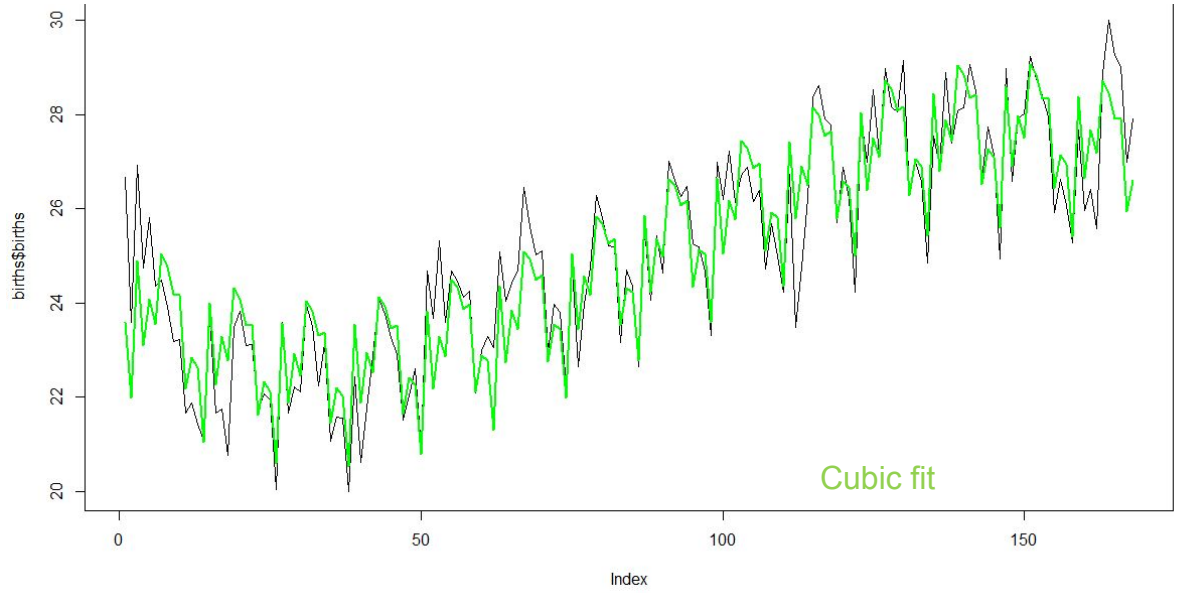

Data Editor

File Edit Help

| | births | time | var3 | var4 | var5 | var6 |
|---|---|---|---|---|---|---|
| 1 | 26.663 | 1 | | | | |
| 2 | 23.598 | 2 | | | | |
| 3 | 26.931 | 3 | | | | |
| 4 | 24.74 | 4 | | | | |
| 5 | 25.806 | 5 | | | | |
| 6 | 24.364 | 6 | | | | |
| 7 | 24.477 | 7 | | | | |
| 8 | 23.901 | 8 | | | | |
| 9 | 23.175 | 9 | | | | |
| 10 | 23.227 | 10 | | | | |
| 11 | 21.672 | 11 | | | | |
| 12 | 21.87 | 12 | | | | |
| 13 | 21.439 | 13 | | | | |
| 14 | 21.089 | 14 | | | | |
| 15 | 23.709 | 15 | | | | |
| 16 | 21.669 | 16 | | | | |
| 17 | 21.752 | 17 | | | | |
| 18 | 20.761 | 18 | | | | |
| 19 | 23.479 | 19 | | | | |

# Seasonal Regression Models - Births




Data Editor

File Edit Help

| | births | time | seasonal | var4 | var5 |
|---|---|---|---|---|---|
| 1 | 26.663 | 1 | 1 | | |
| 2 | 23.598 | 2 | 2 | | |
| 3 | 26.931 | 3 | 3 | | |
| 4 | 24.74 | 4 | 4 | | |
| 5 | 25.806 | 5 | 5 | | |
| 6 | 24.364 | 6 | 6 | | |
| 7 | 24.477 | 7 | 7 | | |
| 8 | 23.901 | 8 | 8 | | |
| 9 | 23.175 | 9 | 9 | | |
| 10 | 23.227 | 10 | 10 | | |
| 11 | 21.672 | 11 | 11 | | |
| 12 | 21.87 | 12 | 12 | | |
| 13 | 21.439 | 13 | 1 | | |
| 14 | 21.089 | 14 | 2 | | |
| 15 | 23.709 | 15 | 3 | | |
| 16 | 21.669 | 16 | 4 | | |
| 17 | 21.752 | 17 | 5 | | |
| 18 | 20.761 | 18 | 6 | | |
| 19 | 23.479 | 19 | 7 | | |

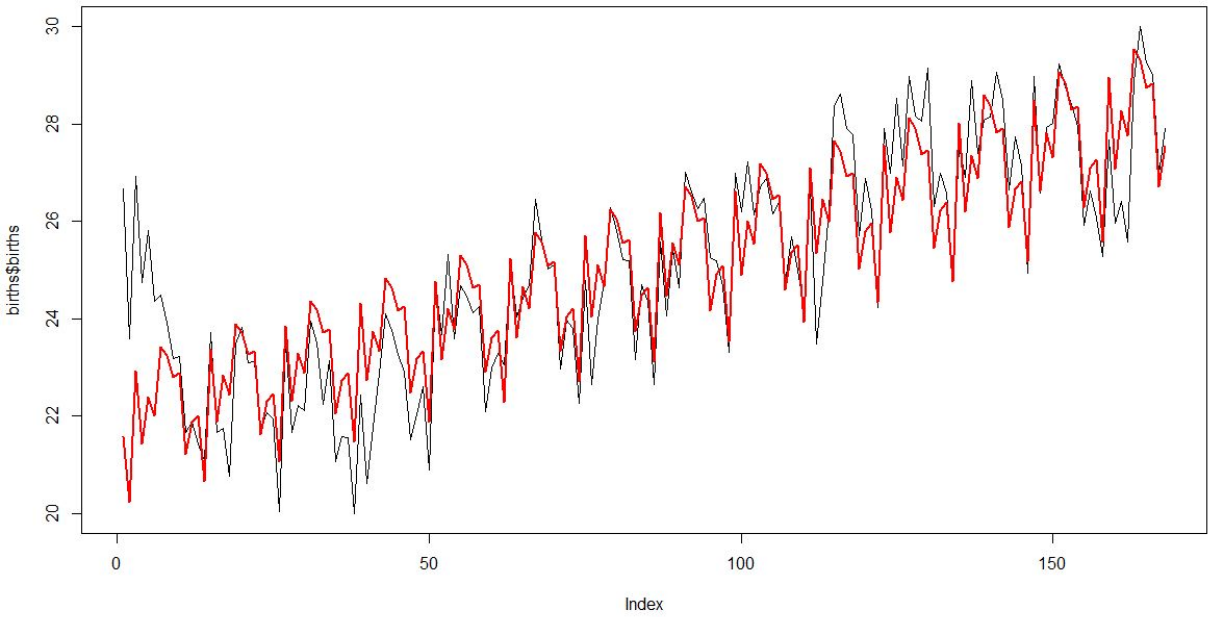
# *forecast* package in R

- The *forecast* package in R contains several time-series forecasting methods.

- *tslm* function captures this break-up of linear model with seasonality

```
>
> tslmfit <- tslm(birthstimeseries ~ poly(trend,3) + season)
>
```

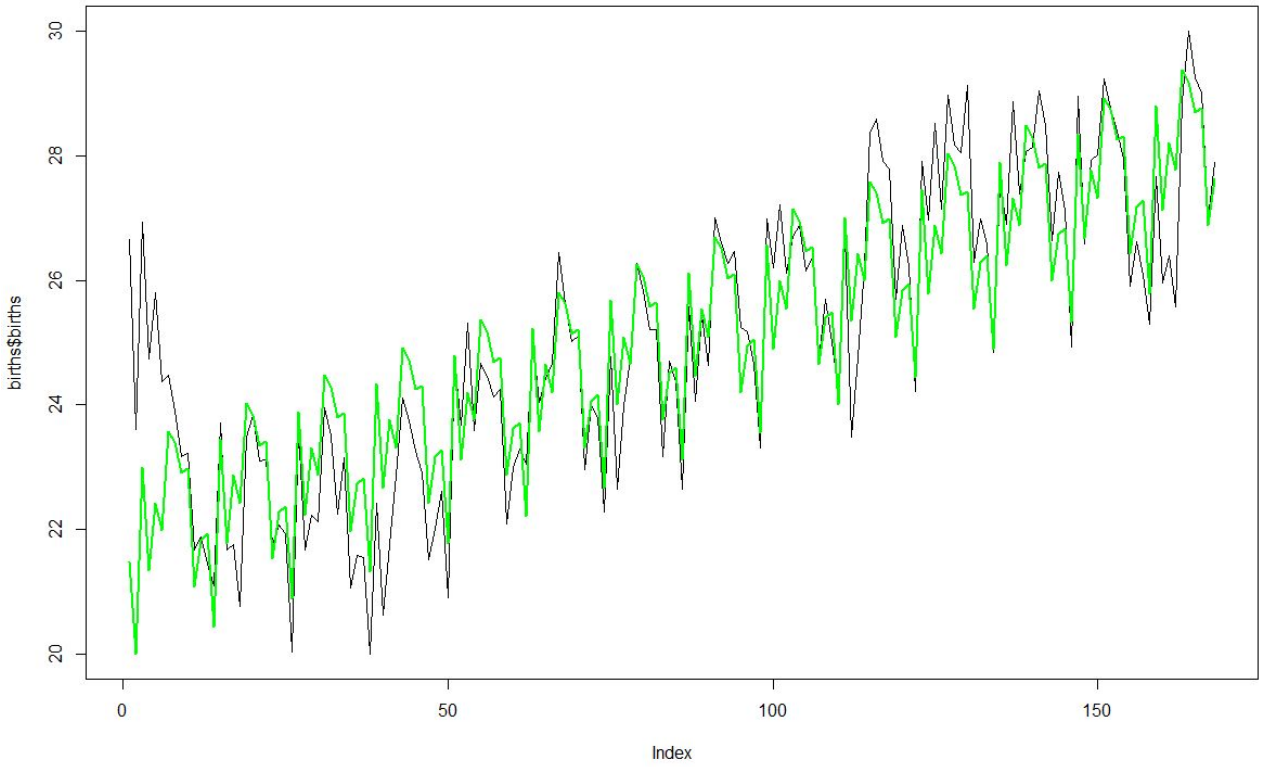
# Seasonality: Multiplicative




| births | time | seasonal | SeasonalFactor |
|---|---|---|---|
| 26.663 | 1 | 1 | 1.2143042 |
| 23.598 | 2 | 2 | 1.0729008 |
| 26.931 | 3 | 3 | 1.2223736 |
| 24.740 | 4 | 4 | 1.1210359 |
| 25.806 | 5 | 5 | 1.1673742 |
| 24.364 | 6 | 6 | 1.1002941 |
| 24.477 | 7 | 7 | 1.1035459 |
| 23.901 | 8 | 8 | 1.0757752 |
| 23.175 | 9 | 9 | 1.0413570 |
| 23.227 | 10 | 10 | 1.0419543 |
| 21.672 | 11 | 11 | 0.9705802 |
| 21.870 | 12 | 12 | 0.9778208 |
| 21.439 | 13 | 1 | 0.9569611 |
| 21.089 | 14 | 2 | 0.9397800 |
| 23.709 | 15 | 3 | 1.0547879 |
| 21.669 | 16 | 4 | 0.9624398 |
| 21.752 | 17 | 5 | 0.9645349 |

```
> births$SeasonalFactor <- births$births/predict(lm1)
> seasonalAdustFactor <- tapply(births$SeasonalFactor,
+                       births$seasonal, mean)
> birthspr <- predict(lm1)*rep(seasonalAdustFactor,14)
> plot(births$births, type="l")
> points(births$time, birthspr,   type="l", col="red", lwd=2)
```

# Seasonality: Additive



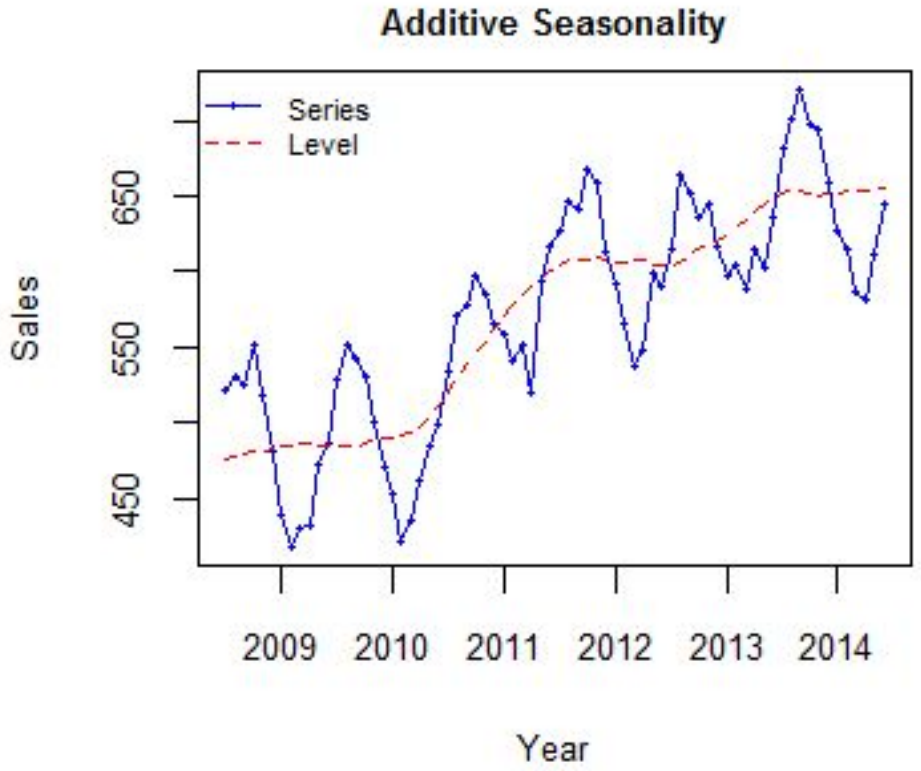

Data Editor

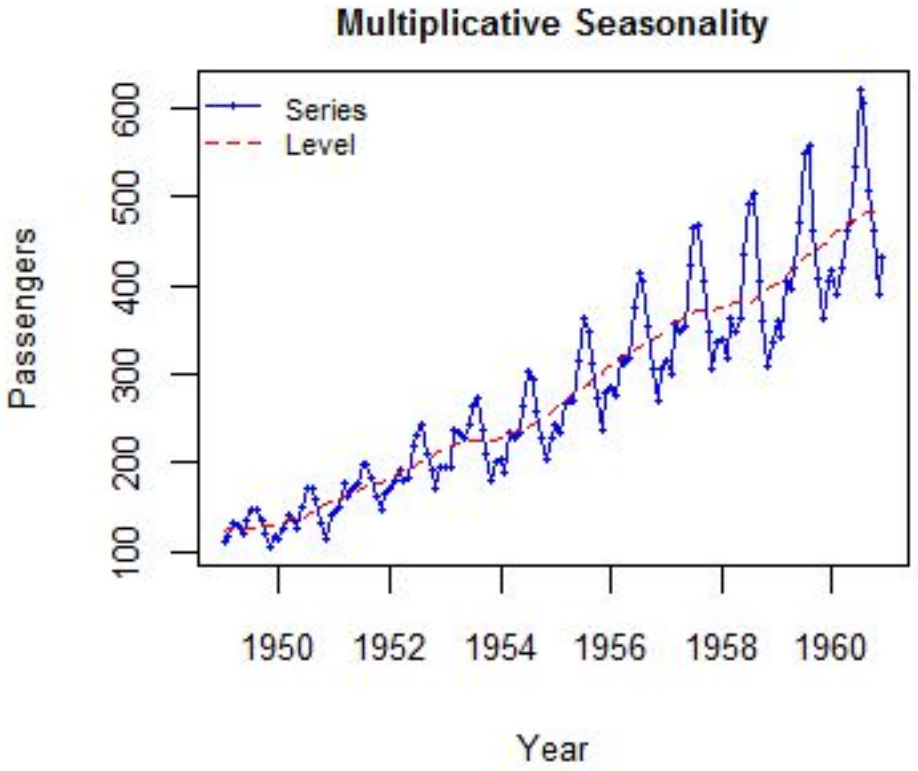
File Edit Help

| | births | time | seasonal | mae |
|---|---|---|---|---|
| 1 | 26.663 | 1 | 1 | 4.70557 |
| 2 | 23.598 | 2 | 2 | 1.603422 |
| 3 | 26.931 | 3 | 3 | 4.899274 |
| 4 | 24.74 | 4 | 4 | 2.671125 |
| 5 | 25.806 | 5 | 5 | 3.699977 |
| 6 | 24.364 | 6 | 6 | 2.220829 |
| 7 | 24.477 | 7 | 7 | 2.29668 |
| 8 | 23.901 | 8 | 8 | 1.683532 |
| 9 | 23.175 | 9 | 9 | 0.920384 |
| 10 | 23.227 | 10 | 10 | 0.9352357 |
| 11 | 21.672 | 11 | 11 | -0.6569126 |
| 12 | 21.87 | 12 | 12 | -0.4960608 |
| 13 | 21.439 | 13 | 1 | -0.9642091 |
| 14 | 21.089 | 14 | 2 | -1.351357 |
| 15 | 23.709 | 15 | 3 | 1.231494 |
| 16 | 21.669 | 16 | 4 | -0.8456539 |
| 17 | 21.752 | 17 | 5 | -0.7998021 |
| 18 | 20.761 | 18 | 6 | -1.82795 |
| 19 | 23.479 | 19 | 7 | 0.8529014 |

```
> births$mae <- births$births-predict(lm1)
> seasonalAdd <- tapply(births$mae,
+                       births$seasonal, mean)
> birthspr <- predict(lm1)+rep(seasonalAdd,14)
> plot(births$births, type="l")
> points(births$time, birthspr, type="l", col="green", lwd=2)
>
```

# Additive or Multiplicative



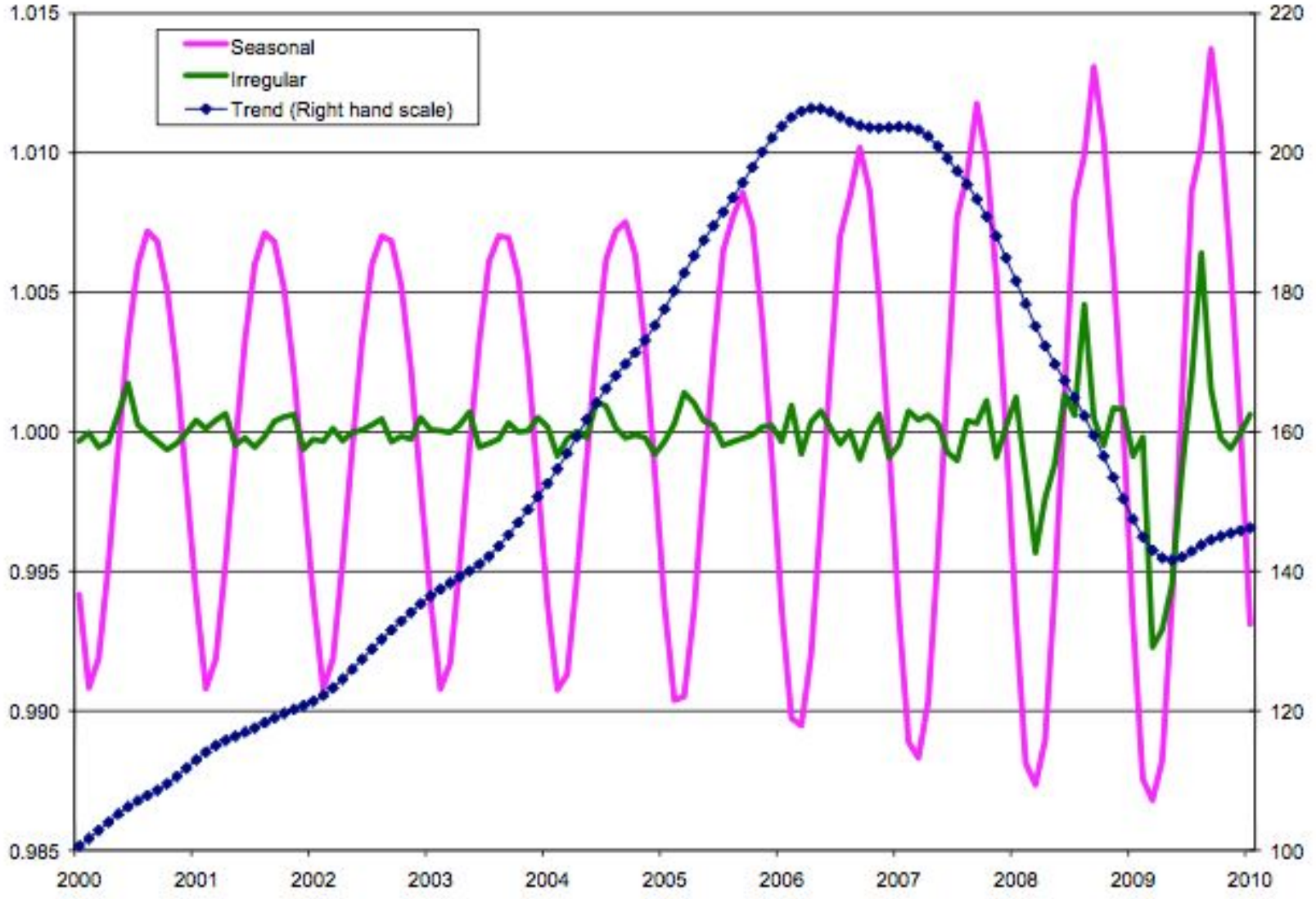

Source: http://www.forsoc.net/2014/11/11/can-you-identify-additive-and-multiplicative-seasonality/

# Components of time series

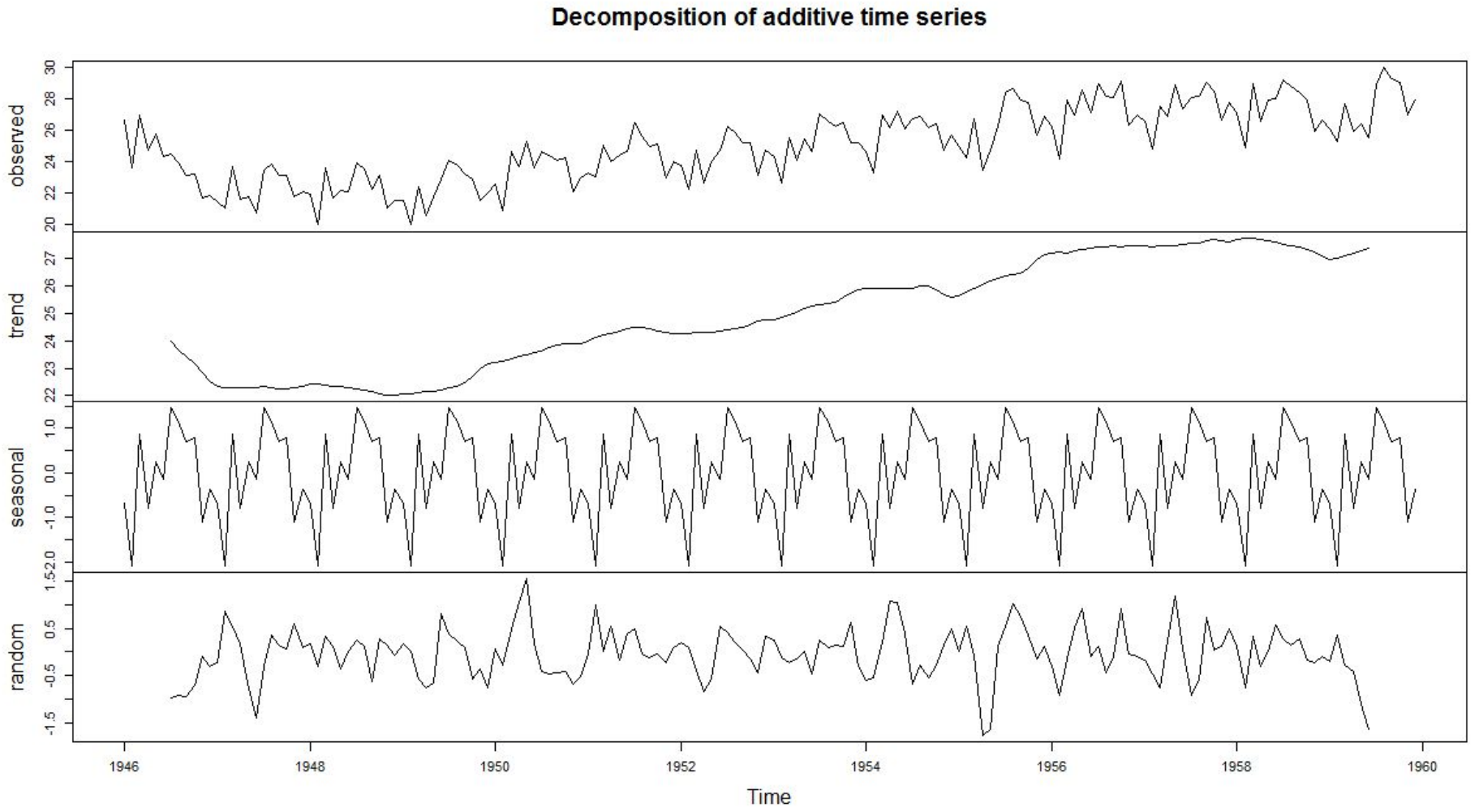
- We use different techniques for time series with different characteristics
  - Trend
  - Seasonal
  - Random stationary

- First we need to identify them

# Trend, Seasonality and Randomness

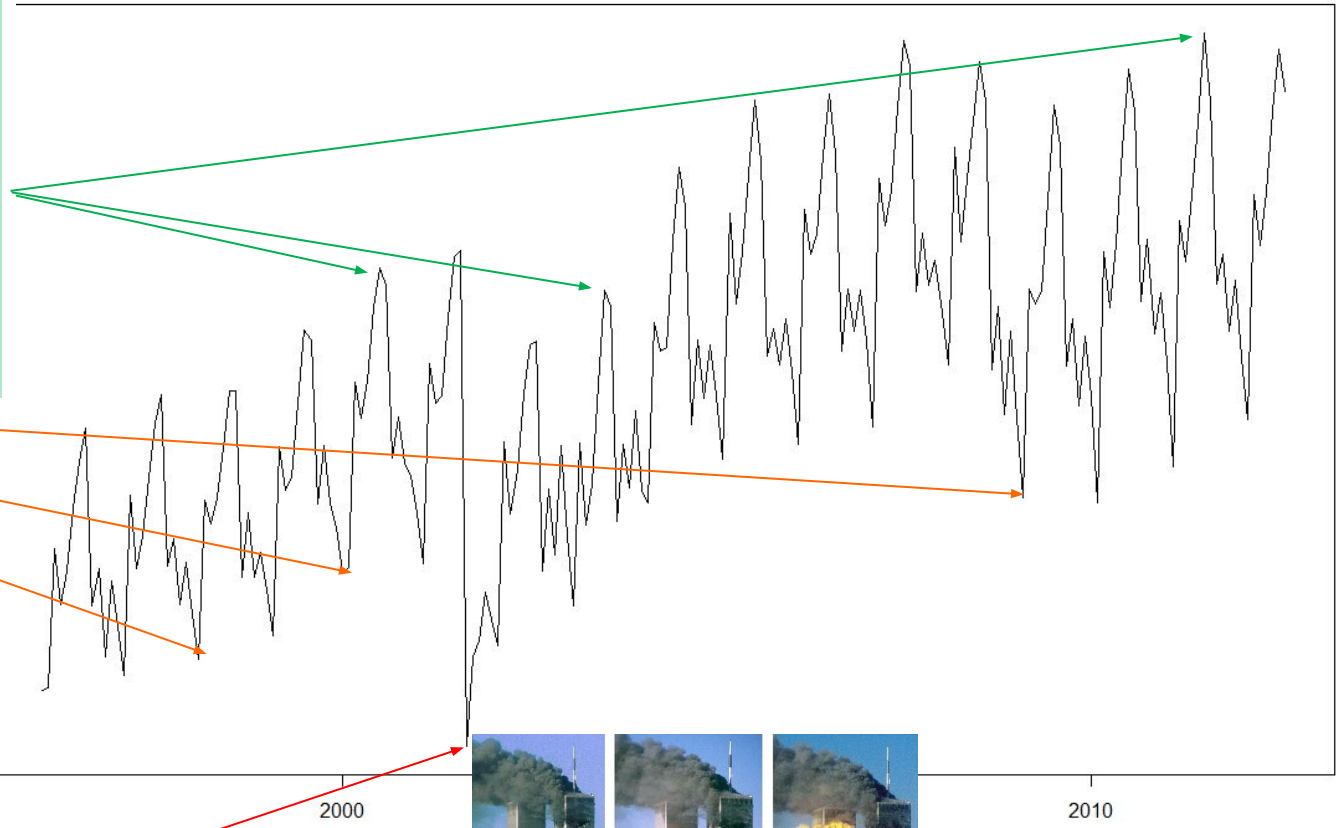


# (Real-world): Decomposing Time Series into the 3 Components – Births in NY



# US Air Carrier Traffic – Revenue Passenger Miles ('000) *RPM*

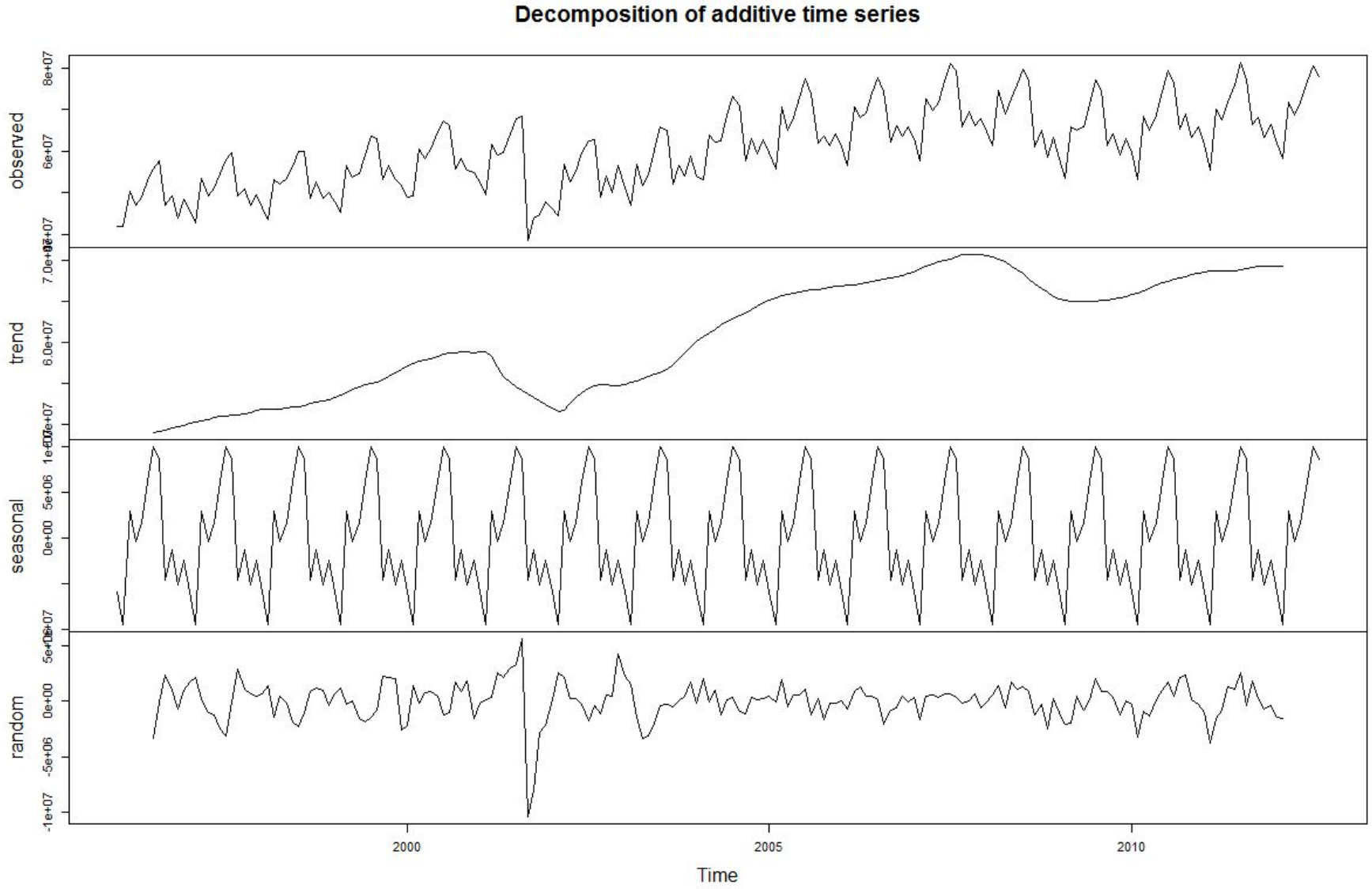
```
> milestimeseries <- ts(miles, frequency = 12, start = c(1996,1))
> milestimeseries
          Jan      Feb      Mar      Apr      May      Jun      Jul
1996 41972194 42054796 50443045 47112397 49118248 52880510 55664750
1997 45850623 42838949 53620994 49282817 51191842 54707221 57995025
1998 46514139 43769273 53361926 51968480 53515798 56460422 59939170
1999 47988560 45241211 56555731 53920855 54674958 59213000 63572248
2000 49045412 49306303 60443541 58286680 60533783 64903295 67346377
2001 52634354 49532578 61575055 59151645 59662416 64353323 67965298
2002 46224031 44615129 56897729 52542164 55116060 59745343 62664511
2003 51197175 47040806 56766580 51857453 54335598 60272900 65962215
2004 53979786 53179693 64035864 62340117 62530704 68866398 73335888
2005 59629608 55795165 70595861 65145552 68268899 72952959 77432998
2006 61035027 56729212 70799794 68120559 69352606 74099239 77798621
2007 63016013 57793832 72700241 69836156 71933109 76926452 80988340
2008 64667106 61504426 74575531 68906882 72725750 76162105 79707545
2009 58373786 53506580 66027341 65166300 65868254 71350227 77136799
2010 59651061 53240066 68307090 64953250 68850904 74474550 79304441
2011 61630362 55391206 70158268 67683558 71711448 76057910 81423215
2012 61940180 58243763 71696039 68669228 71887523 76760759 80499353
          Aug      Sep      Oct      Nov      Dec
1996 57723208 47035464 49263120 43937074 48539606
1997 59715433 49418190 51058879 47056048 49654209
```




Data sources:
http://www.bts.gov/xml/air_traffic/src/index.xml
and
https://datamarket.com/data/set/281x/us-air-carrier-traffic-statistics-revenue-passenger-miles
Last accessed: 31-Mar-2016

| | Aug | Sep | Oct | Nov | Dec |
|---|---|---|---|---|---|
| 1996 | 57723208 | 47035464 | 49263120 | 43937074 | 48539606 |
| 1997 | 59715433 | 49418190 | 51058879 | 47056048 | 49654209 |
| 1998 | 59927214 | 48751280 | 52578217 | 48734375 | 50208641 |
| 1999 | 63003663 | 53131972 | 56653901 | 53215500 | 51746821 |
| 2000 | 66256804 | 55900504 | 58373996 | 55590325 | 54822970 |
| 2001 | 68377080 | 38601868 | 43964788 | 44915764 | 47836501 |
| 2002 | 62944816 | 49096035 | 54019748 | 50106814 | 56656594 |
| 2003 | 64989766 | 52121480 | 56724551 | 54128776 | 58739845 |
| 2004 | 70961522 | 57881042 | 63021142 | 59453943 | 62680310 |

# (Real-world): Decomposing Time Series into the 3 Components – Revenue Passenger Miles (RPM)



# Issues with Regressing on Time

- If there is no trend or if seasonality and fluctuations are more important than trend, then the coefficients behave weirdly

# TIME SERIES: AUTO REGRESSIVE METHODS

$$\hat{y}_{t+1} = f(y_t, y_{t-1}, y_{t-2} \ldots)$$

# Time Series Descriptive Statistics

- In descriptive statistics covered earlier (central tendencies, measures of variability, skewness, kurtosis, distributions, correlations, etc.), the order of observations in the data was of no consequence.

- In time series descriptive statistics, order of observations is of primary importance and so autocorrelations play a vital role in identifying the models and their characteristics.

- Autocorrelation is a metric that allows us to understand the strength of order in the time-series

# Assumptions : Autoregressive models

- Linear relationship between successive values.
- Additive errors
- Normal independent identically distributed (iid) errors
- The process is stationary

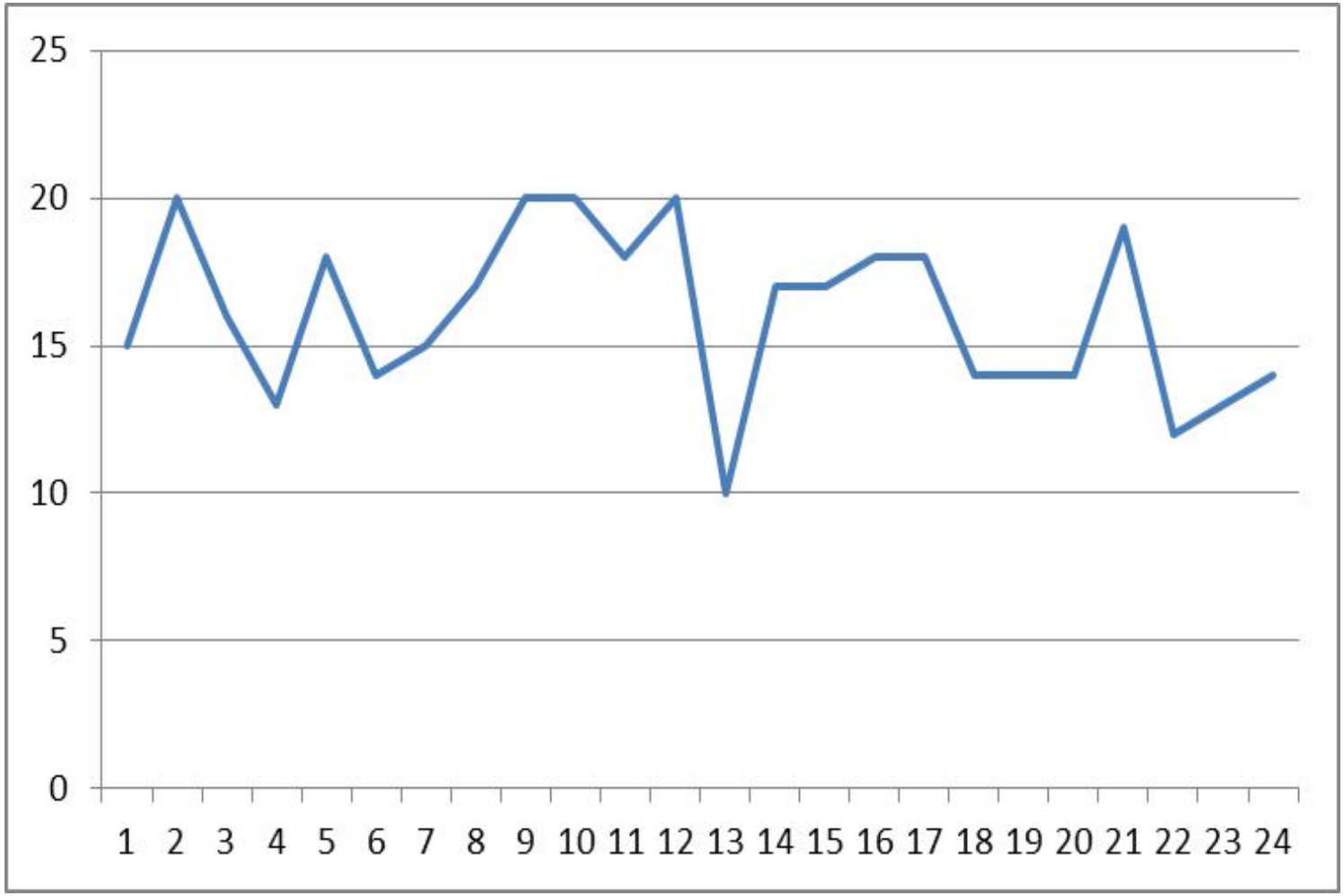
# Stationary and Non-Stationary

- Stationary data has constant statistical properties such as mean, variance, autocorrelation – over time

- If the data is stationary, forecasting is easier!
  - Whether a series is stationary or not can be assessed either through statistical tests such as the Dickey-Fuller test or ACF and PACF plots(explained next).

  - Differencing is a commonly used method to convert non-stationary to stationary (explained later with an example).

# Illustration of a stationary process



# AUTOCORRELATION AND PARTIAL AUTOCORRELATION

# Autocorrelation (ACF) and Partial ACF (PACF)

- ACF: $n^{th}$ lag of ACF is the correlation between a day and $n$ days before that.

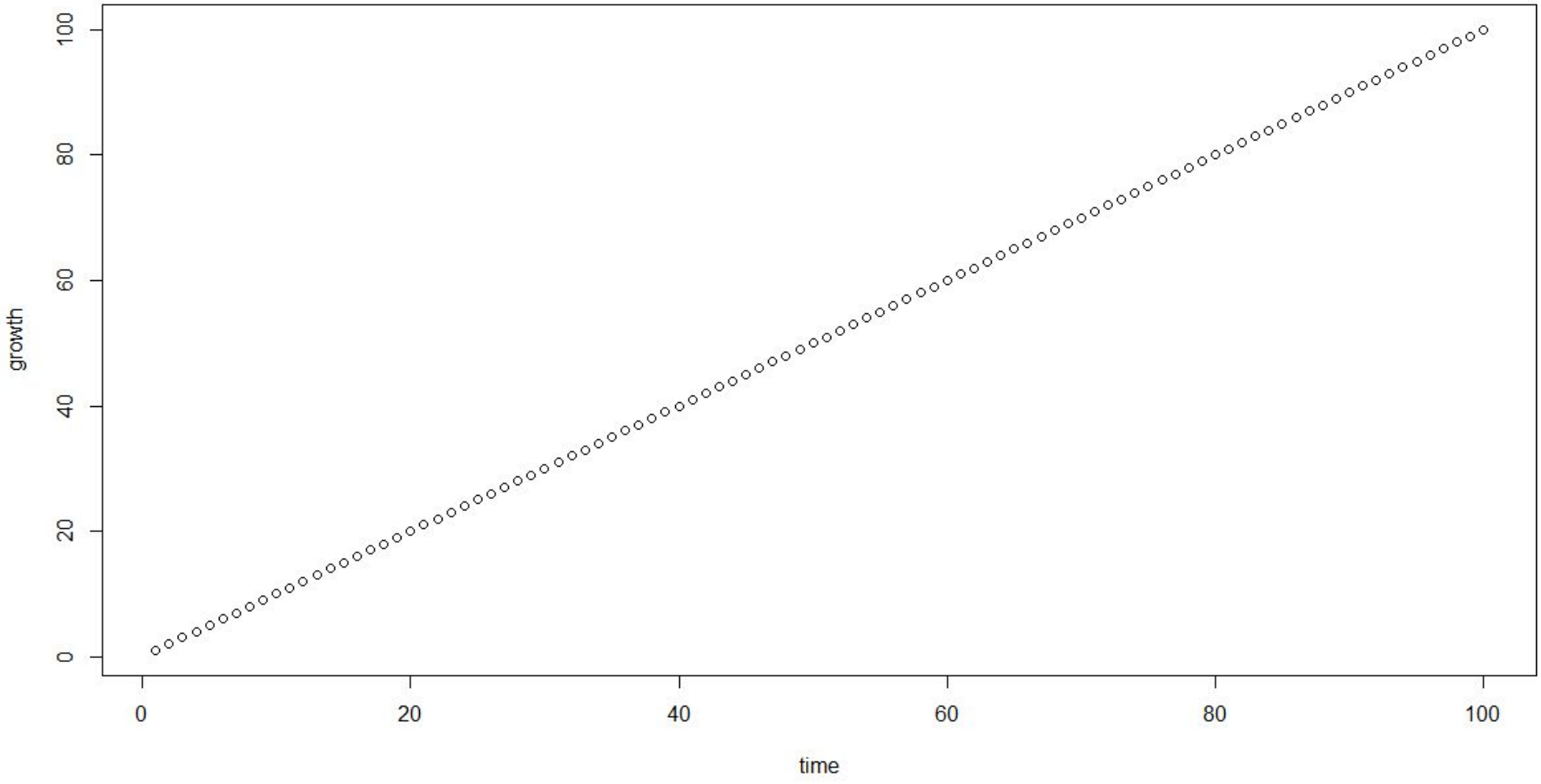
- PACF: The same as ACF with all intermediate correlations removed. It is the $k_{th}$ coefficient of the ordinary least squares regression.

$$[y_t] = \beta_0 + \sum_{i=1}^{k} \beta_i [y_{t-i}] \text{ where}$$

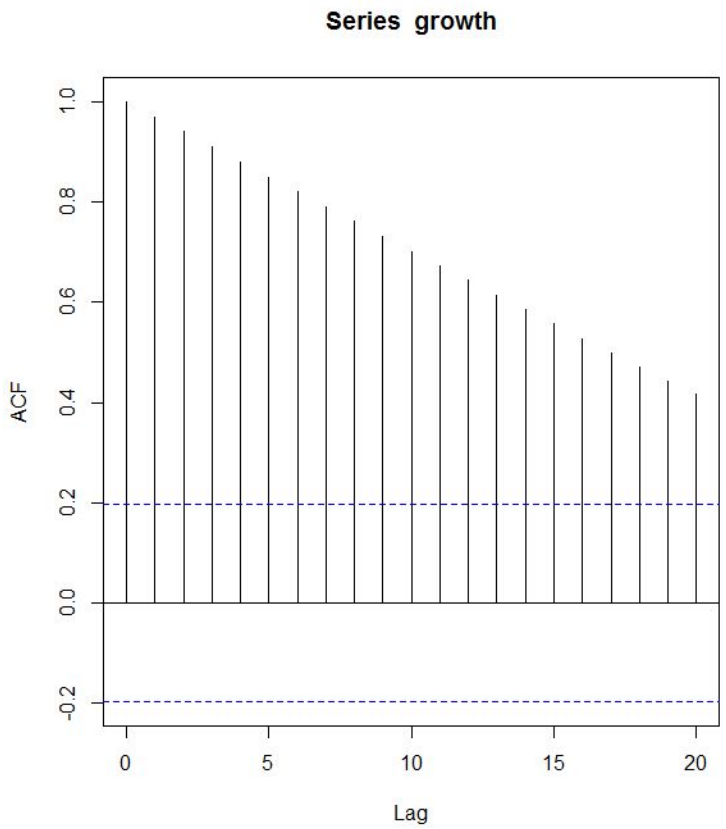
$[y_t]$ is the input time series, $k$ is the lag order and $\beta_i$ is the $i_{th}$ coefficient of the linear multiple regression.

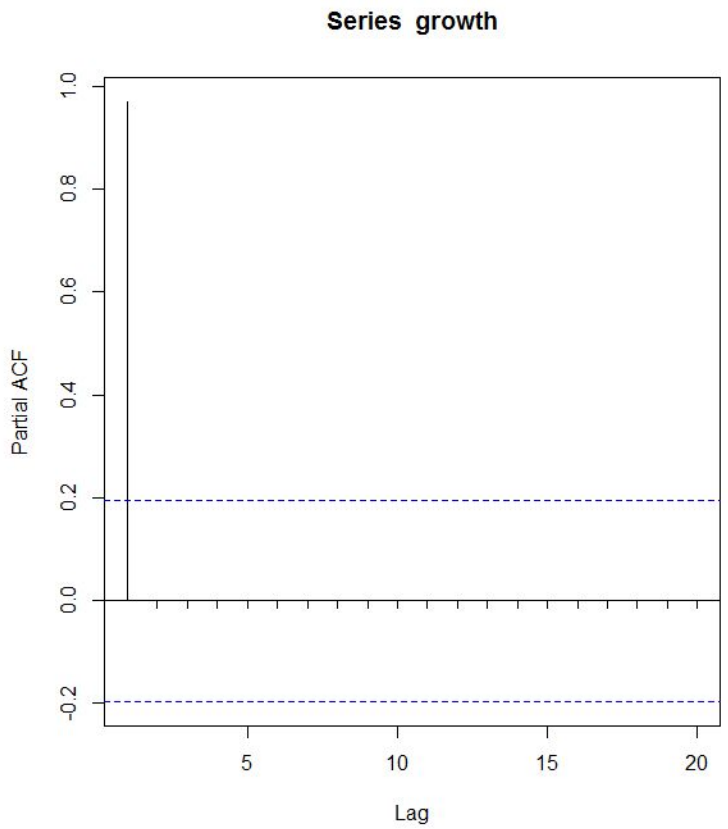
**EXCEL ACTIVITY**

# Idealized Trend



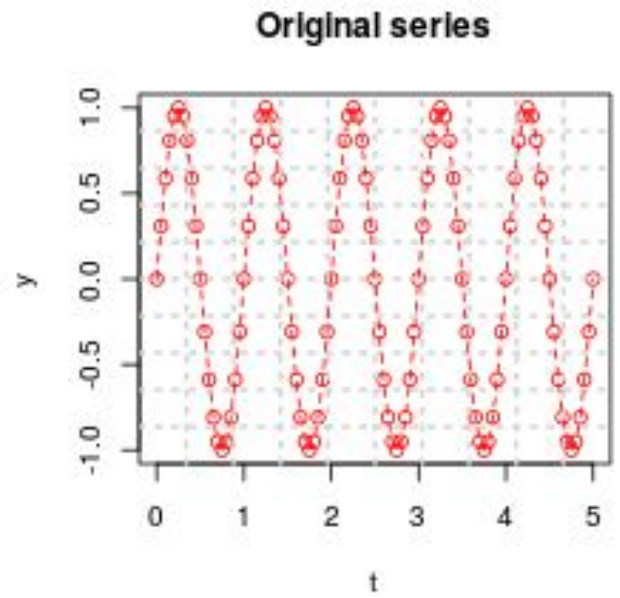
# ACF and PACF – Idealized Trend

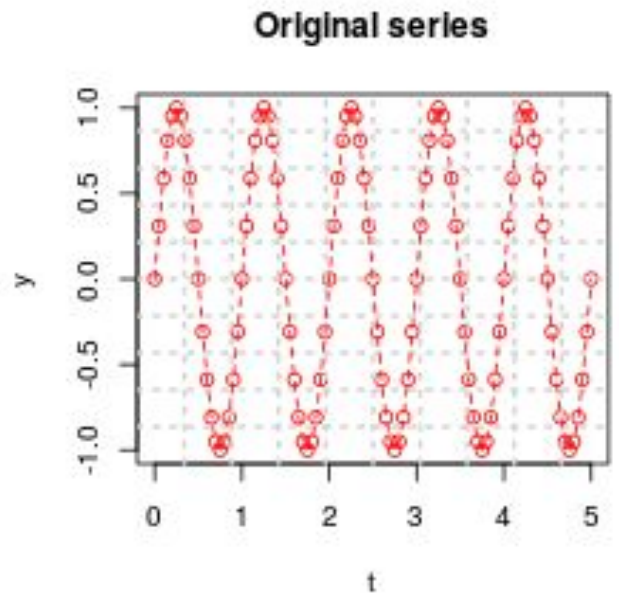


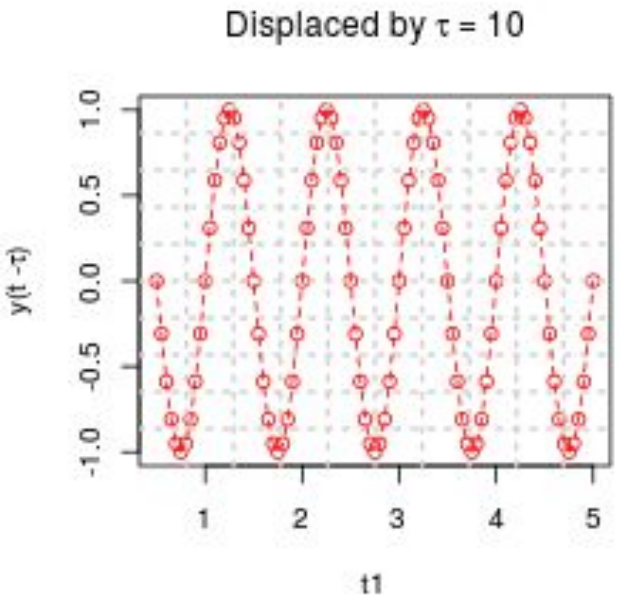



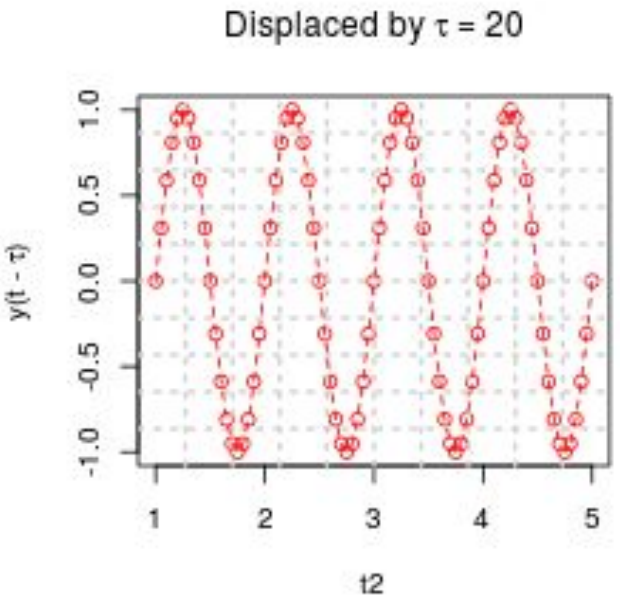

95% CI: $0 \pm \frac{1.96}{\sqrt{n}}$

- ACF is a bar chart of correlation coefficients of the time series and its lags.
- PACF is a plot of the partial correlation coefficients of the time series and its lags.



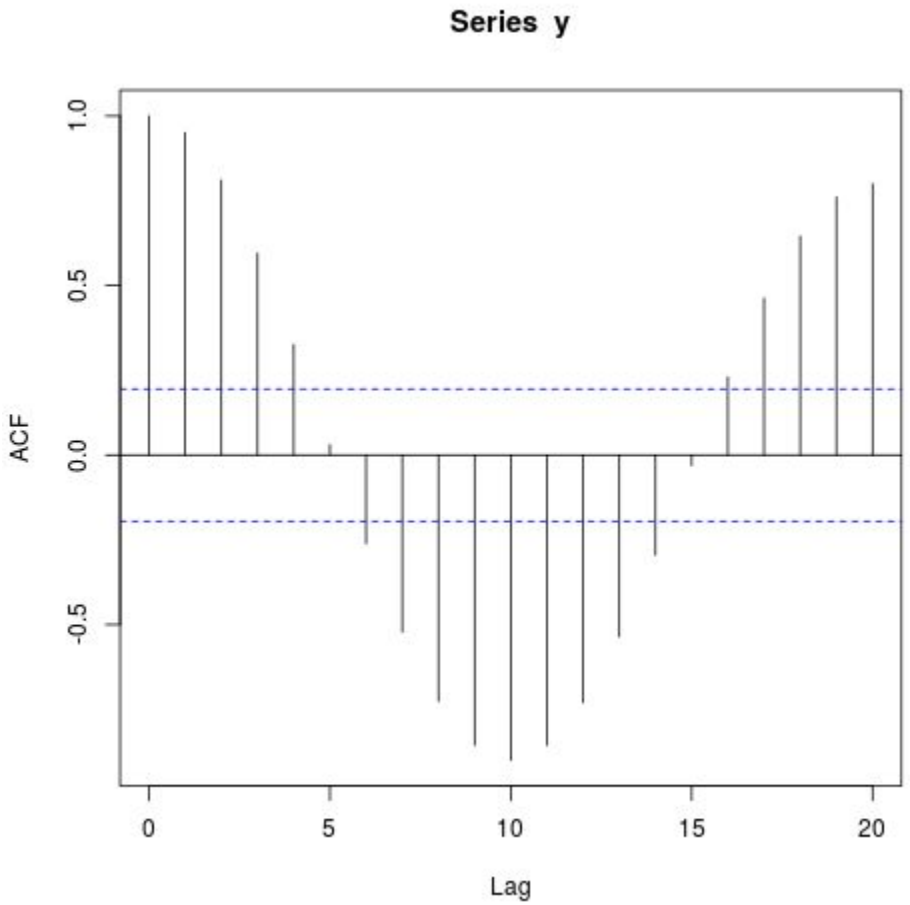



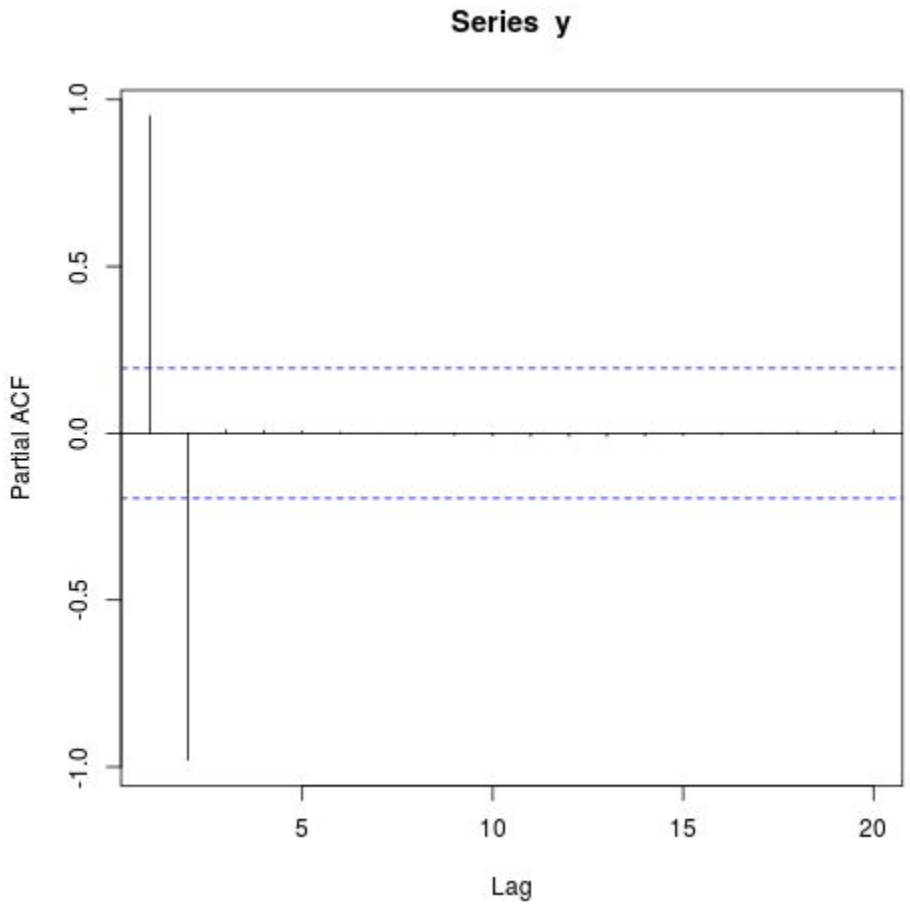




**A periodic time series (sine function)**

# ACF and PACF for sine function

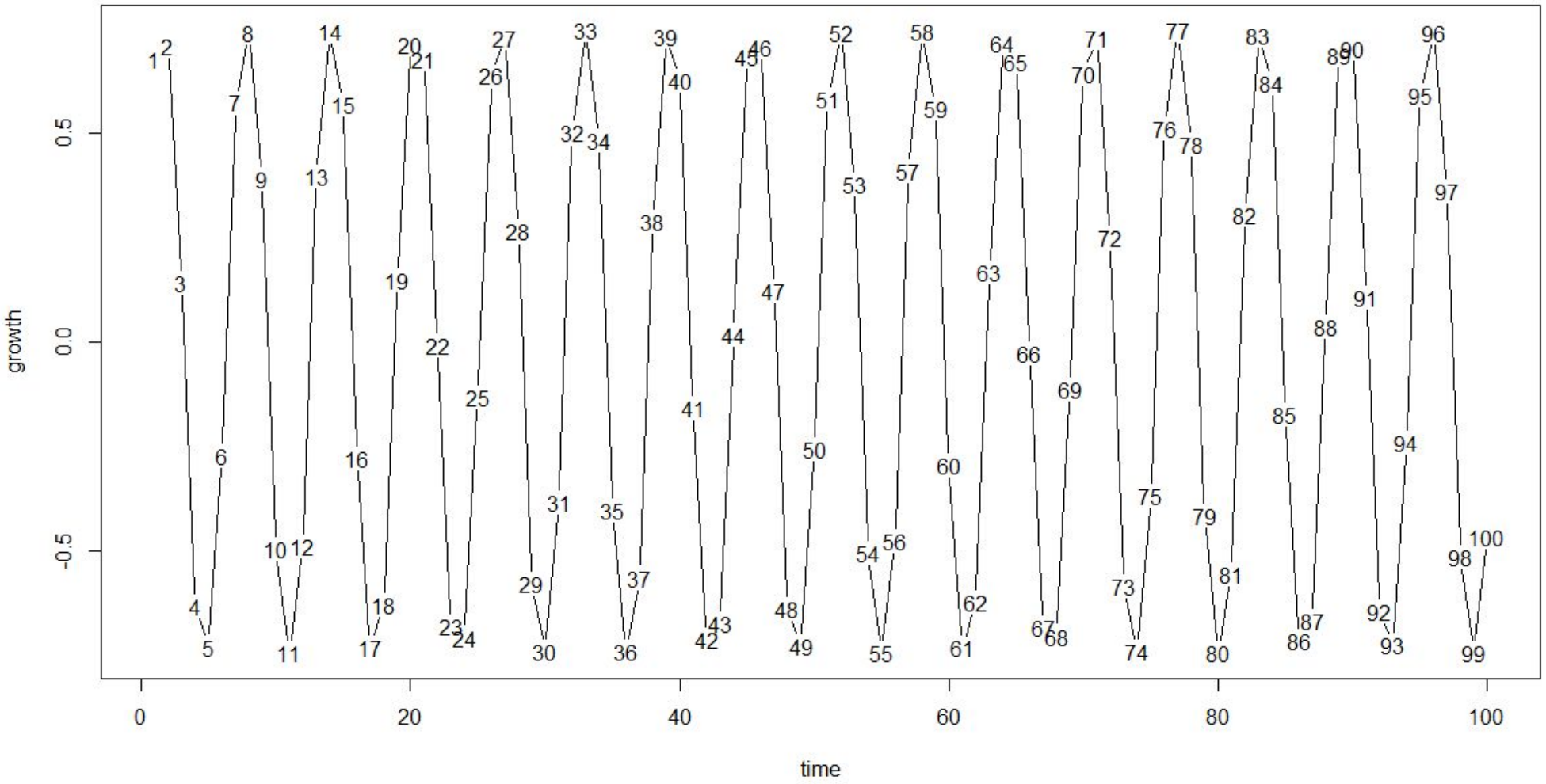



ACF

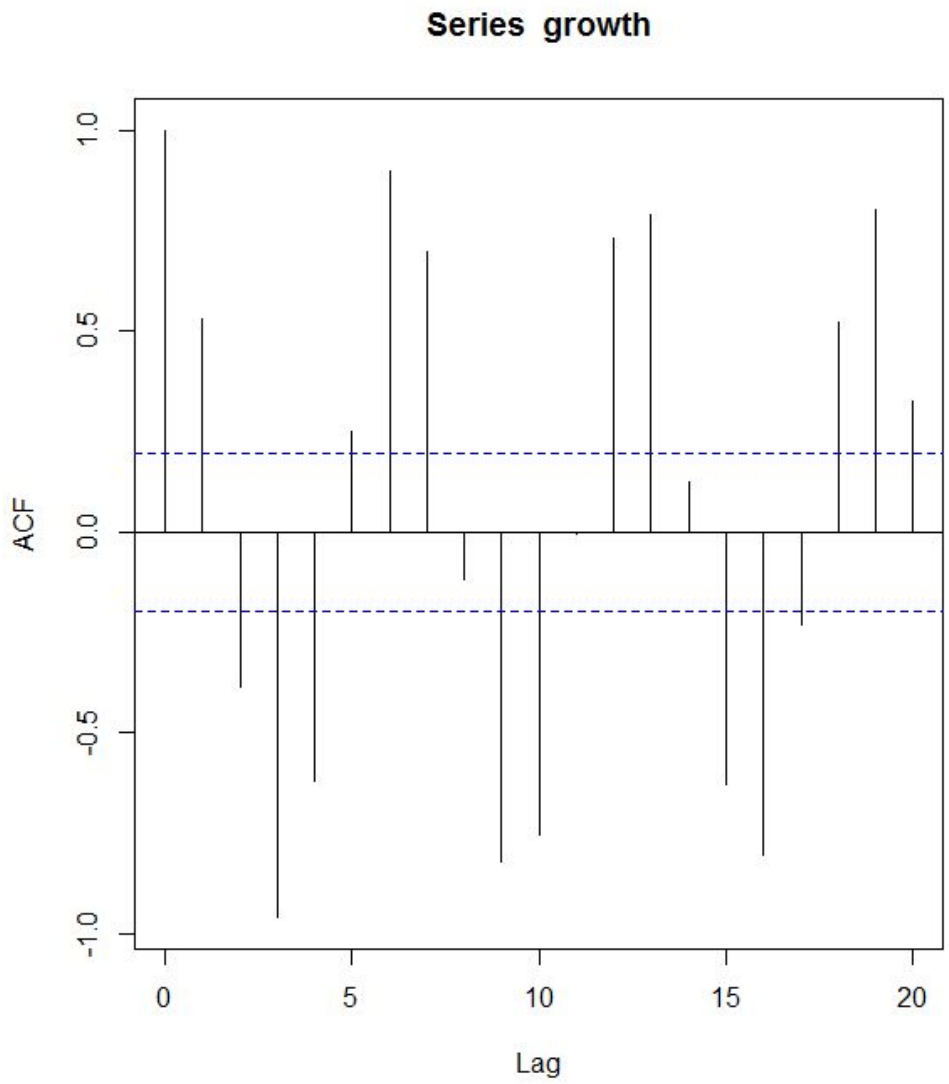


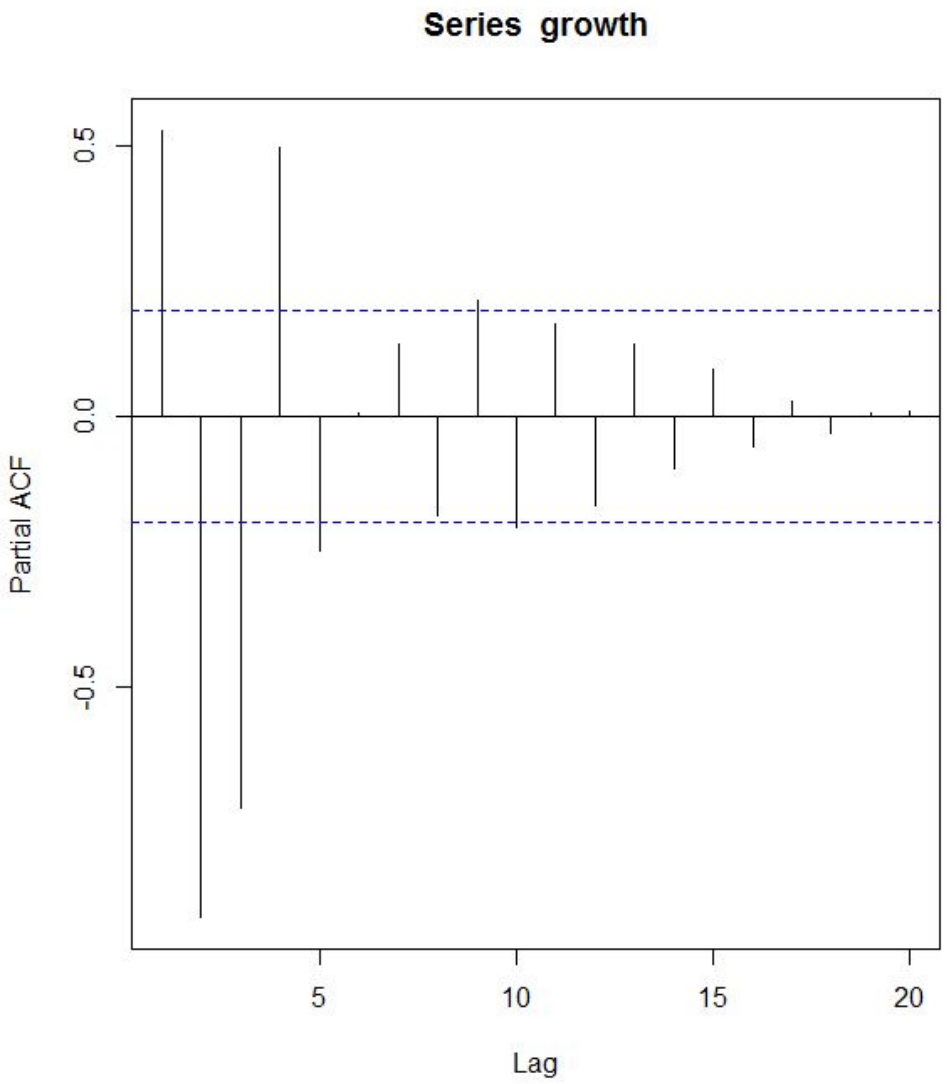

PACF

# ACF and PACF – Idealized Seasonality

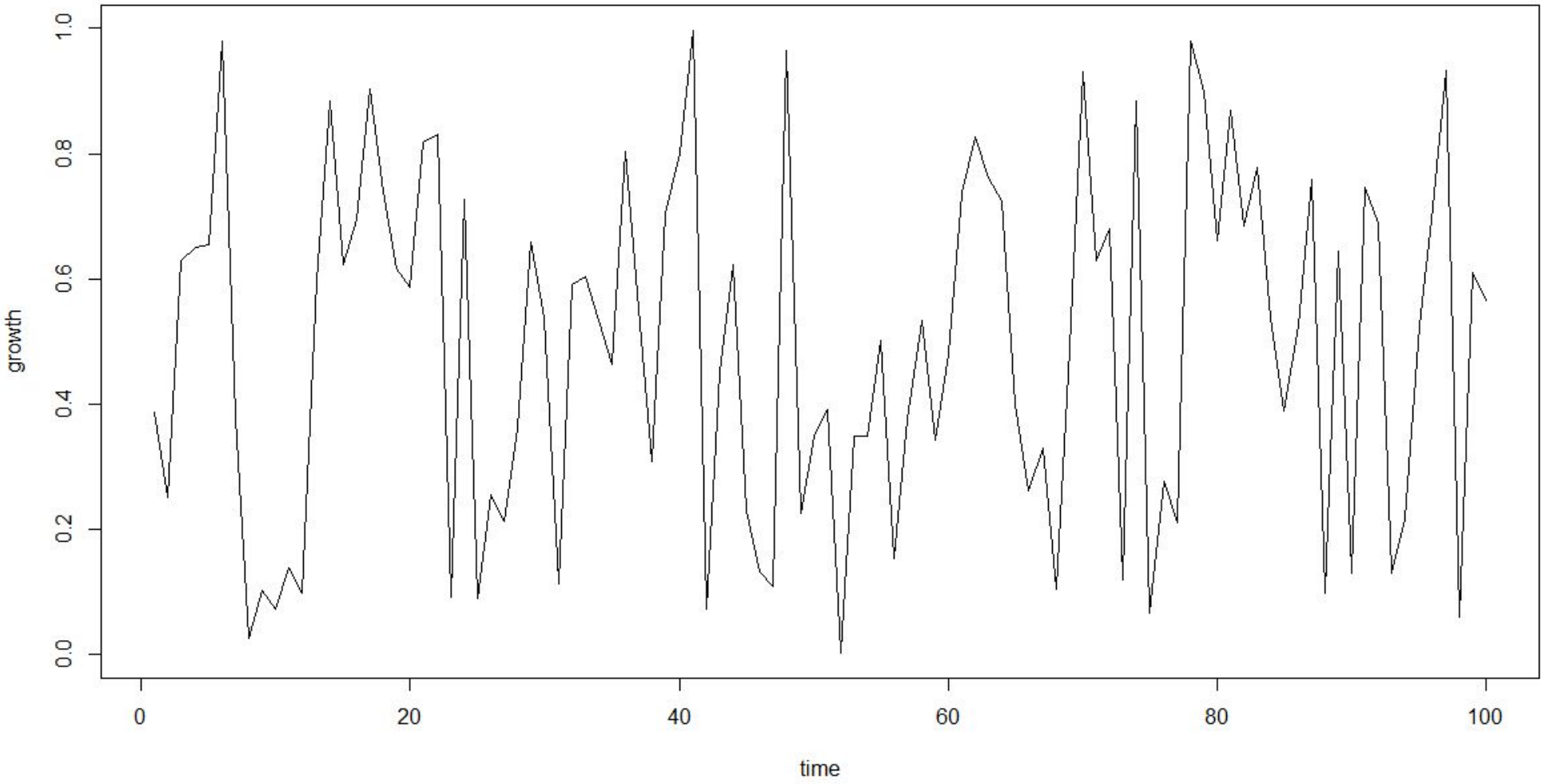


# ACF and PACF – Idealized Seasonality

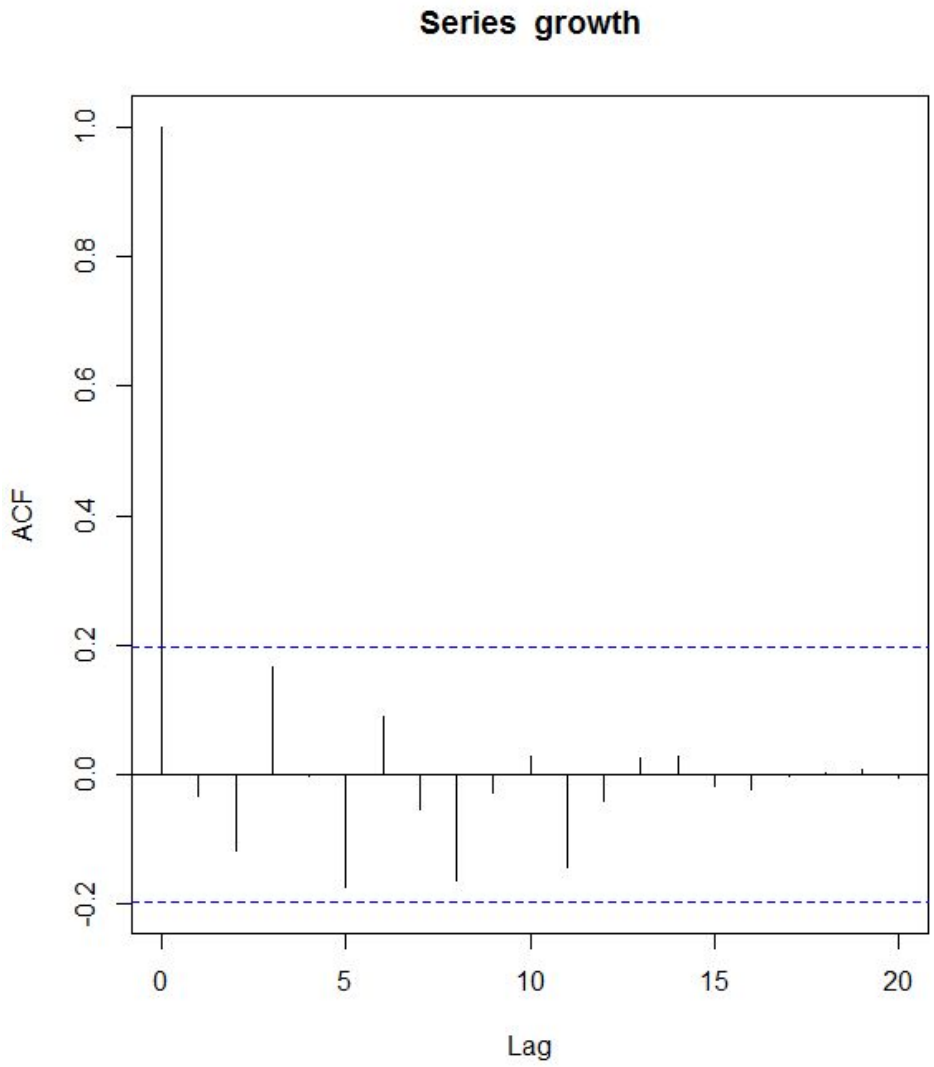


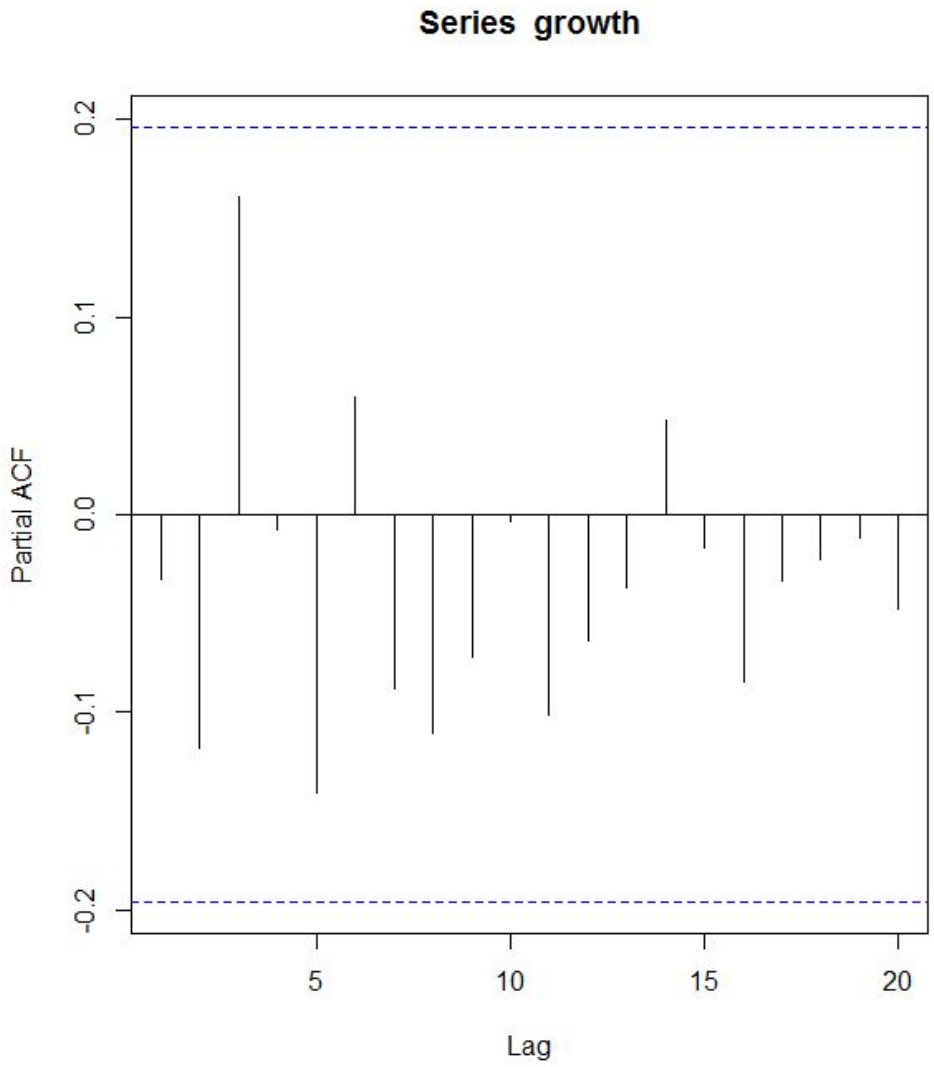



# ACF and PACF – Idealized Randomness



# ACF and PACF – Idealized Randomness

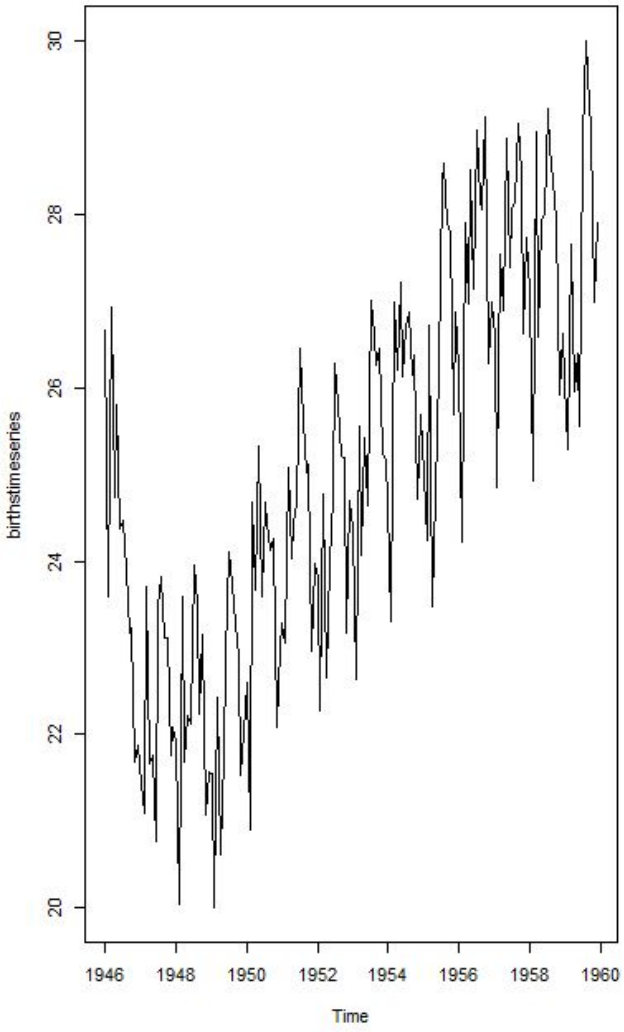


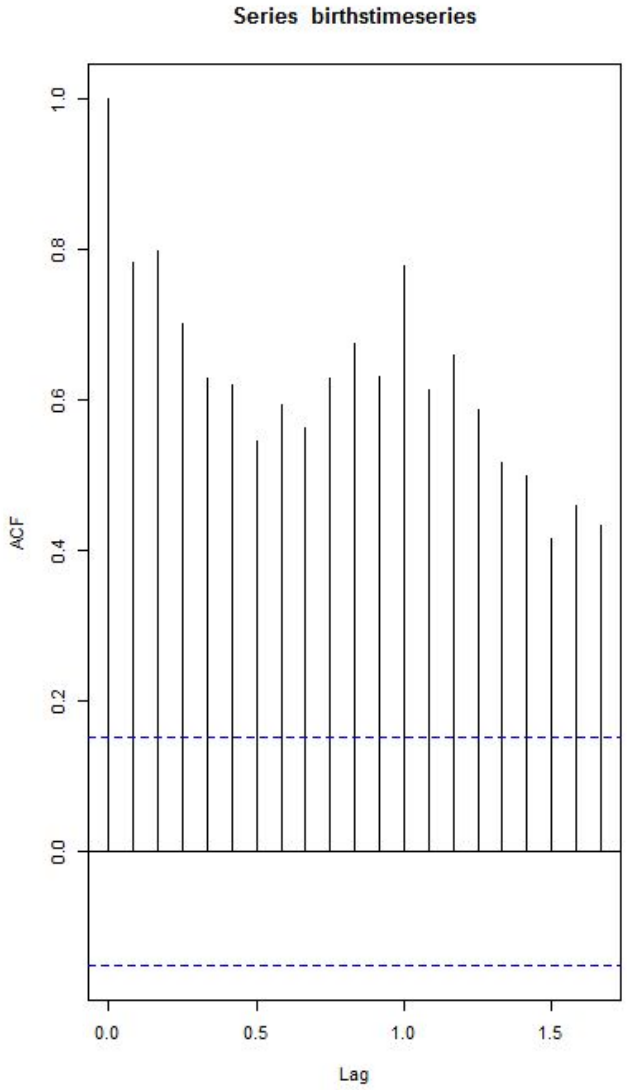



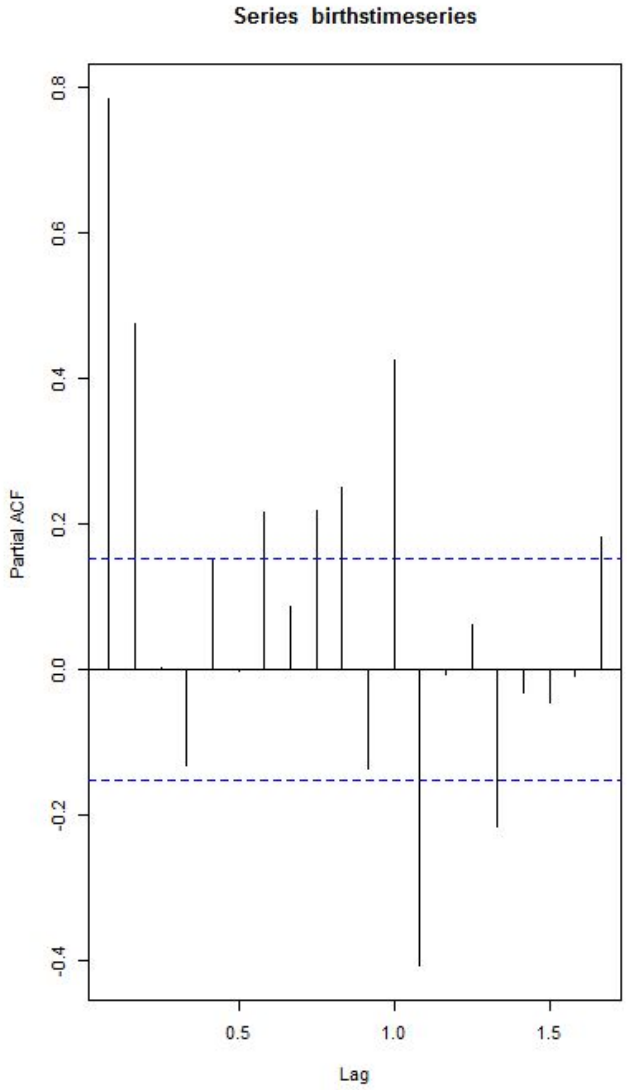
# ACF and PACF – Idealized Trend, Seasonality and Randomness

- Ideal Trend: Decreasing ACF and 1 or 2 lags of PACF
- Ideal Seasonality: Cyclicality in ACF and a few lags of PACF with some positive and some negative
- Ideal Random: A spike may or may not be present; even if present, magnitude will be small

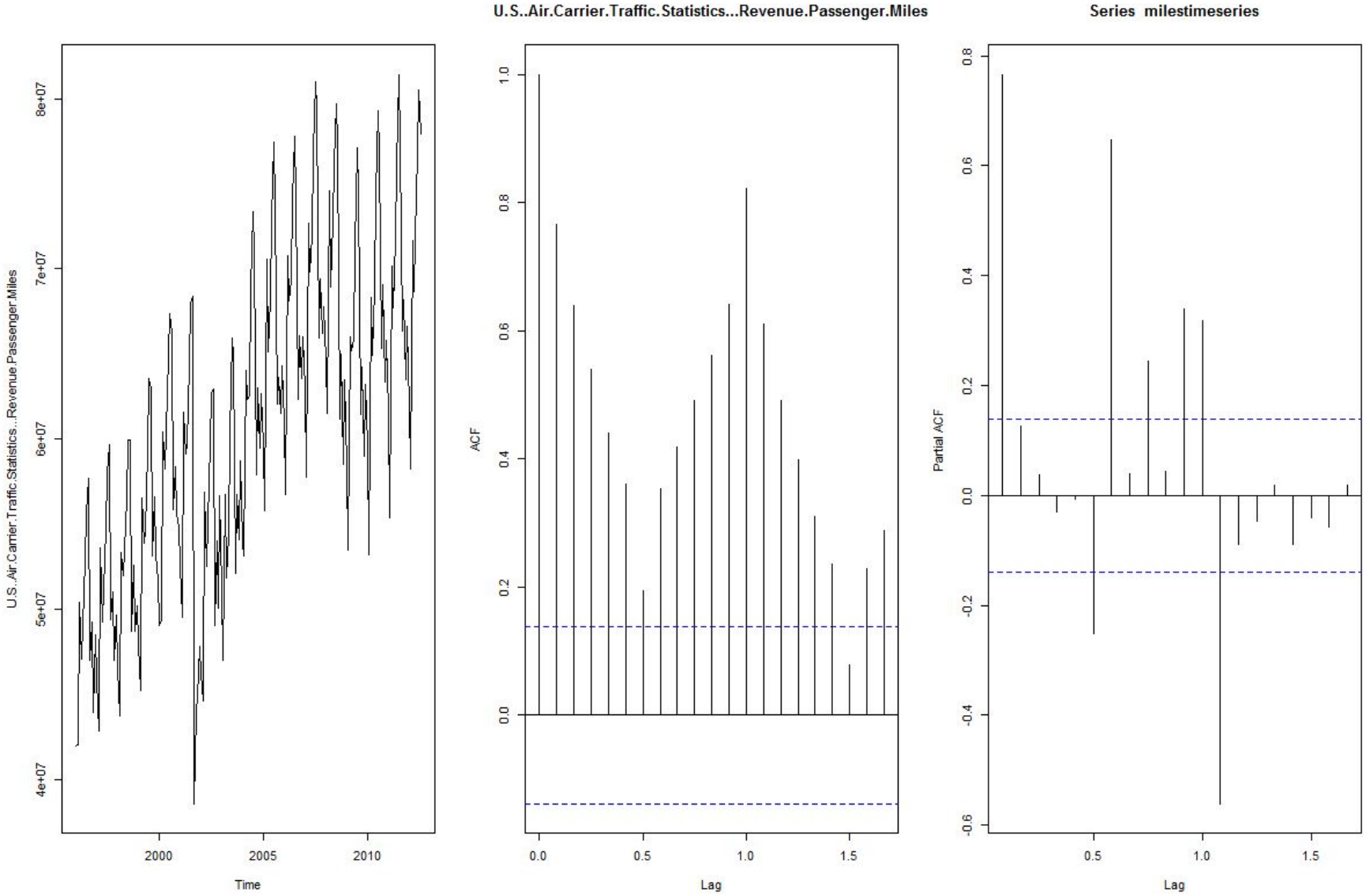
# ACF and PACF (Real-world): Births in NY



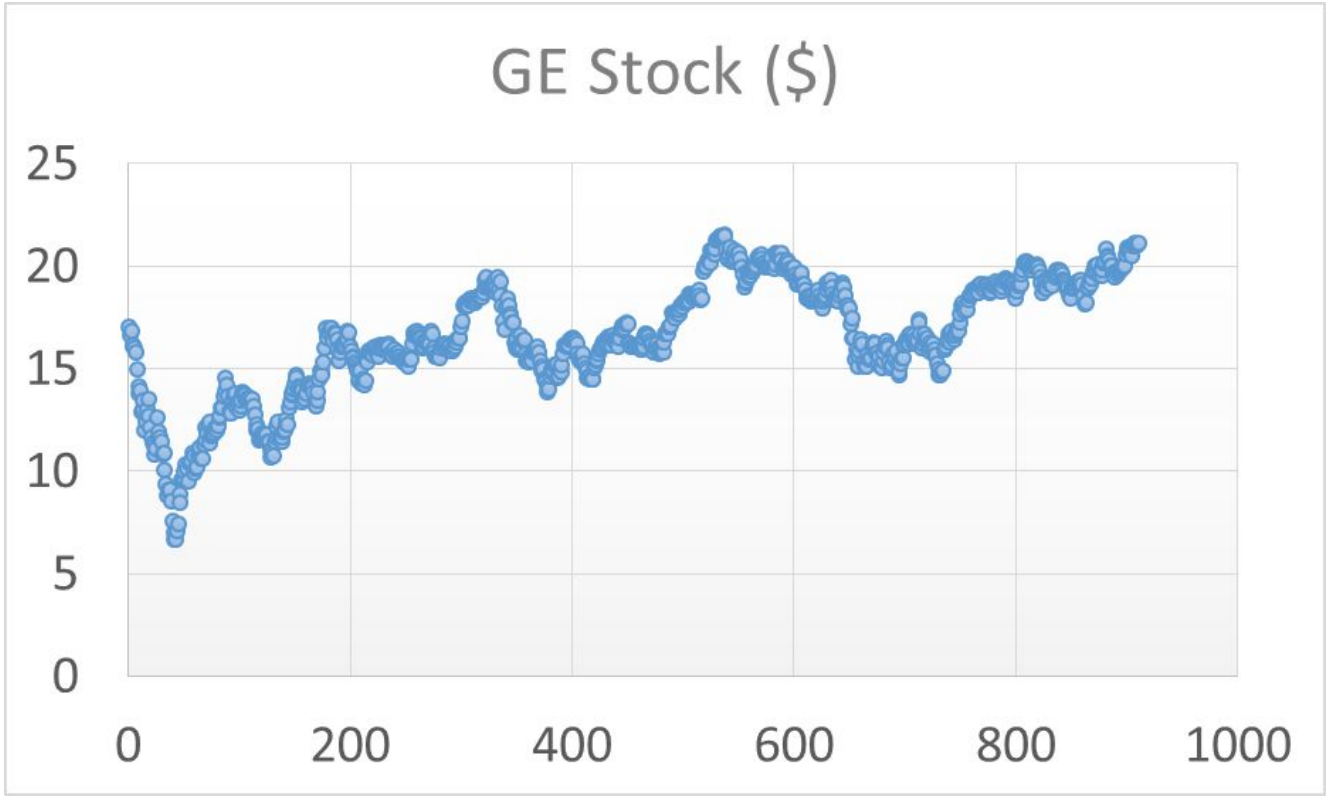



# ACF and PACF (Real world) : Revenue Passenger Miles

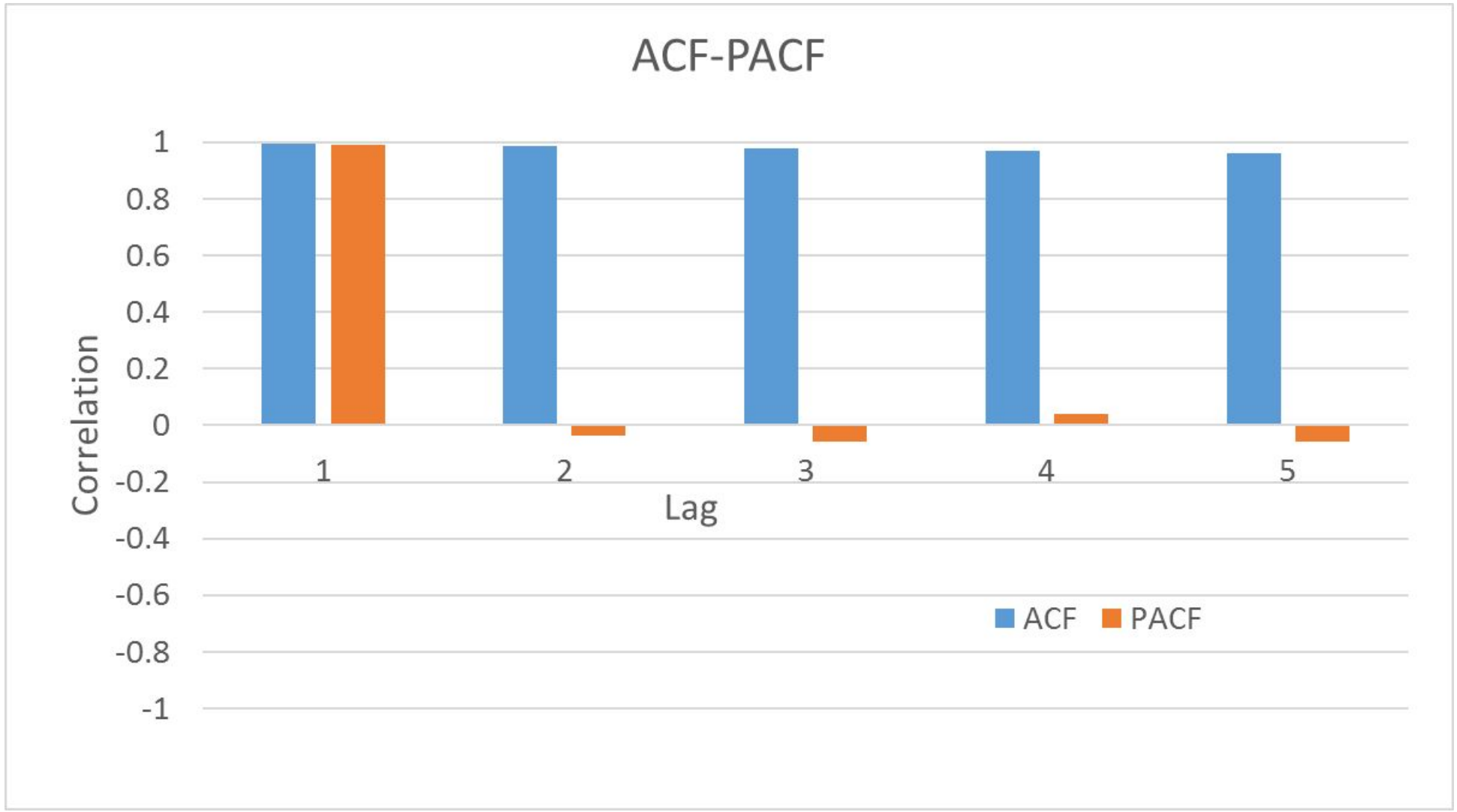


# An example : Stock prices



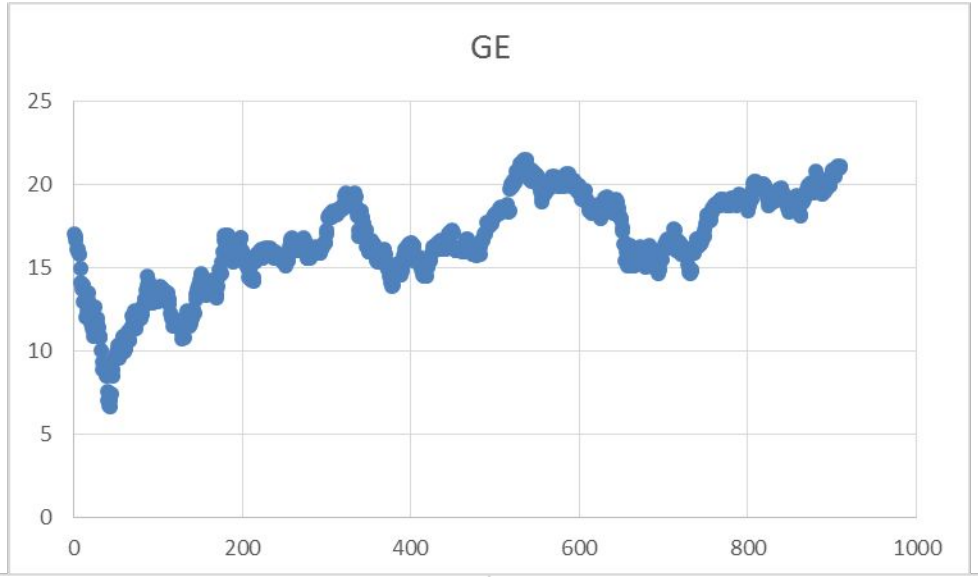

- Series does not appear stationary.
- But do not guess, instead inspect ACF and PACF plots.

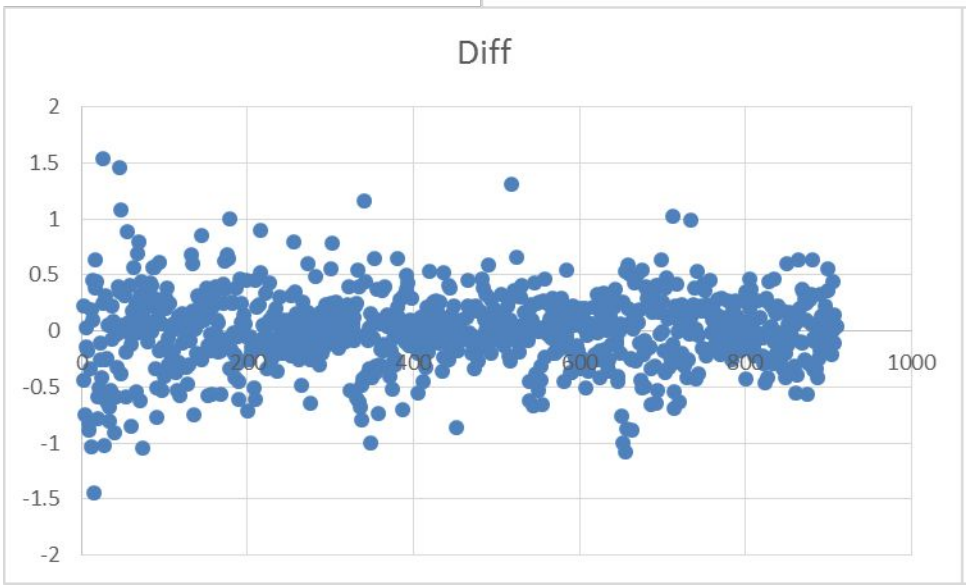
# Autocorrelation (ACF) and Partial ACF (PACF)



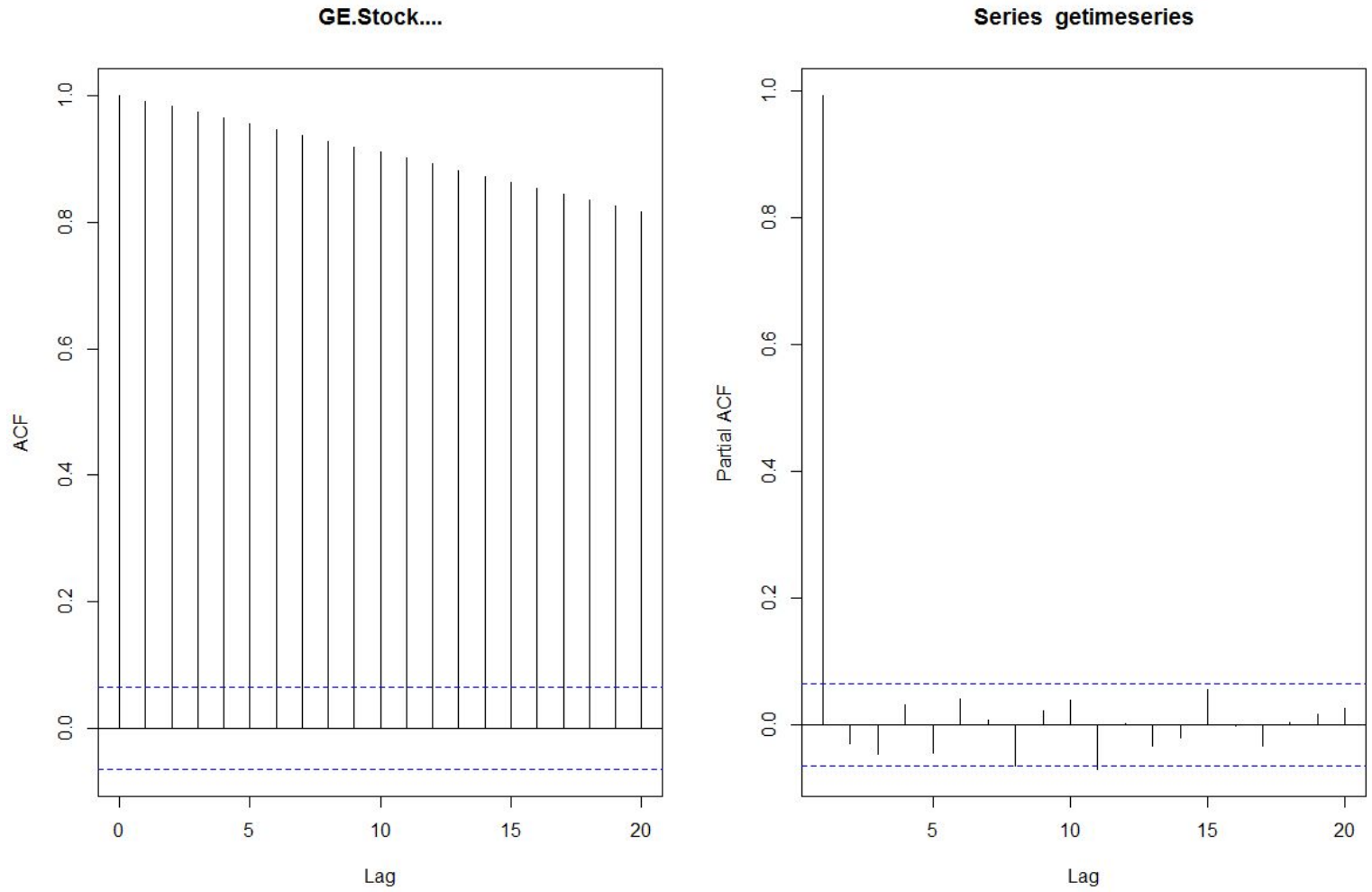

See the attached file 01Correlations.xlsx

# Removing Trend from Data



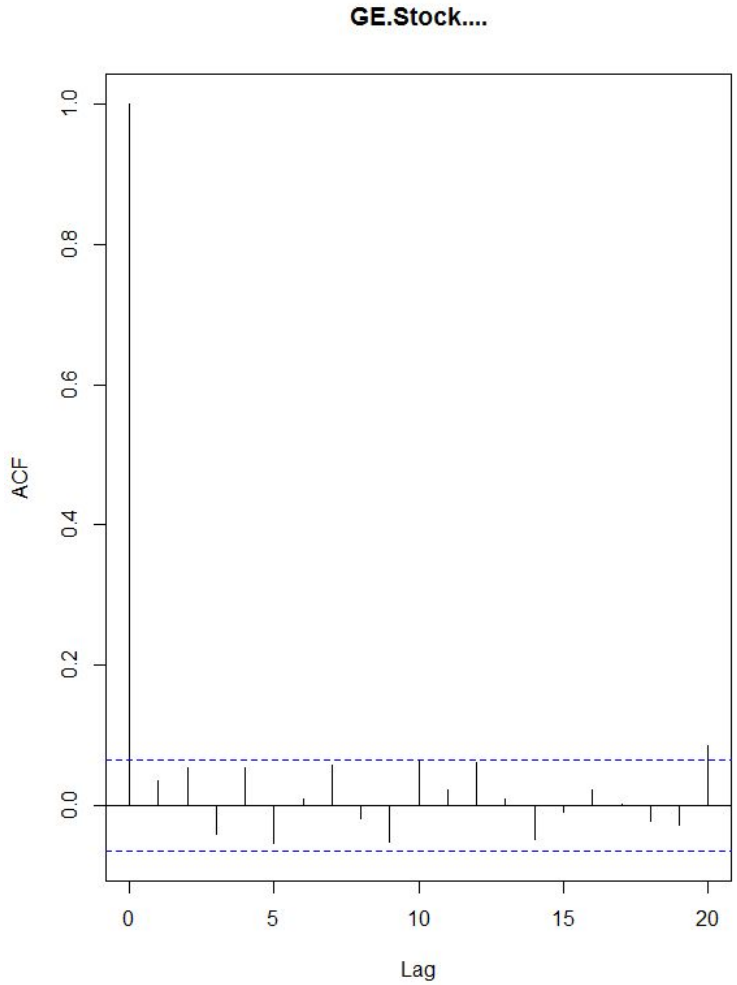



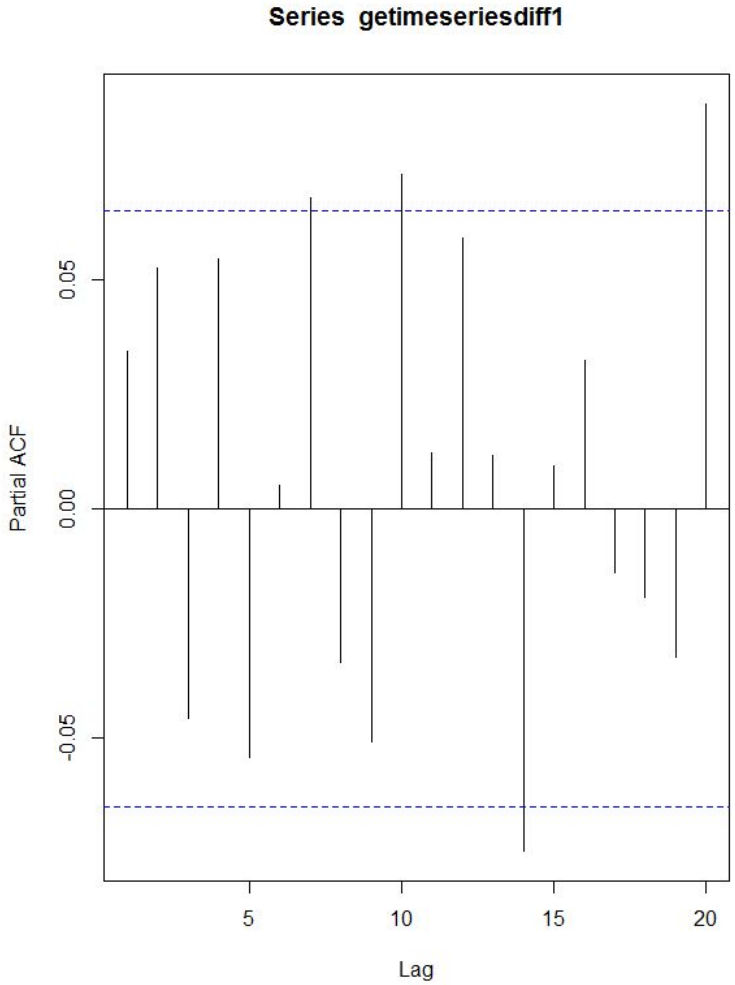

# ACF and PACF of Stationary and Non-Stationary




Price of GE stock is highly correlated with the previous day's value.

# ACF and PACF of Stationary and Non-Stationary




Daily changes in GE stock price are essentially random.

# ACF and PACF of Stationary and Non-Stationary

- Non-stationary series have an ACF that remains significant for half a dozen or more lags, rather than quickly declining to zero.

- You must difference such a series until it is stationary before you can identify the process.

# Moving Averages

- Work best for stationary processes.

- Commonly used moving average methods
  - Simple Moving Average (SMA)
  - Weighted Moving Average (WMA)
  - Exponential Moving Average (EMA) also called Exponential Smoothing

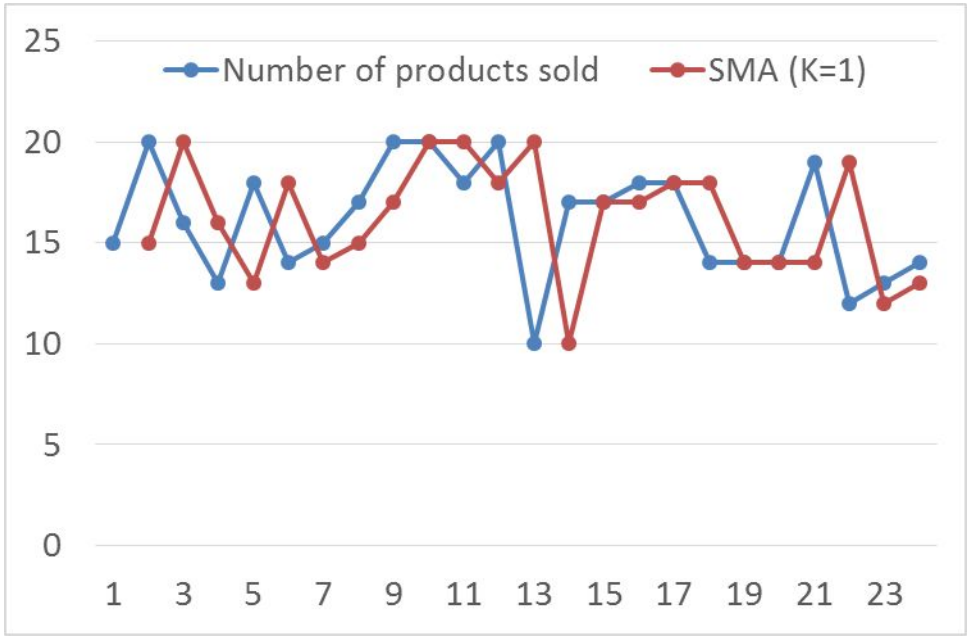
**Excel file example on number of products sold weekly**

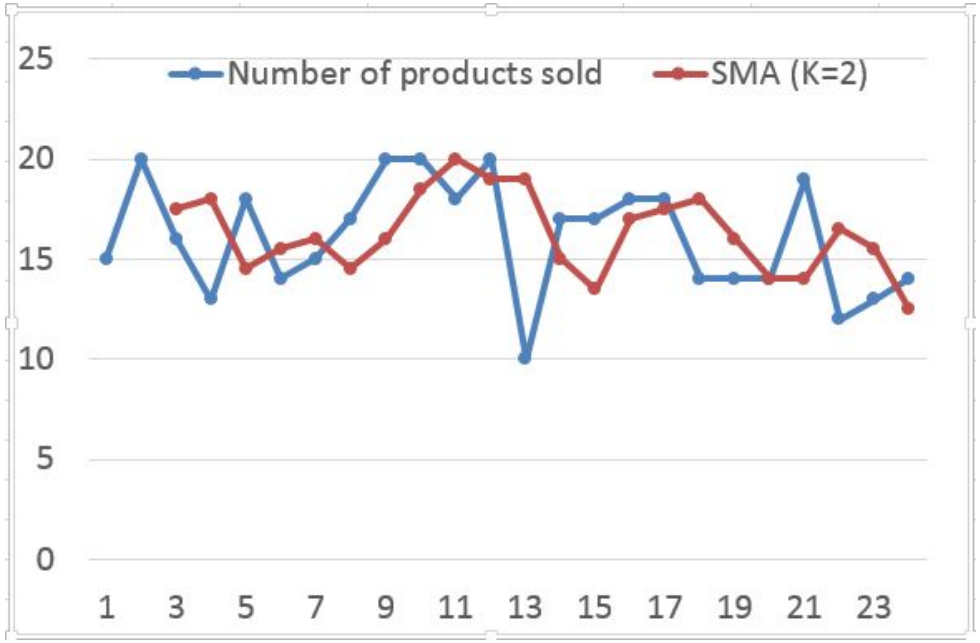
# Stationary Model : Simple Moving Averages

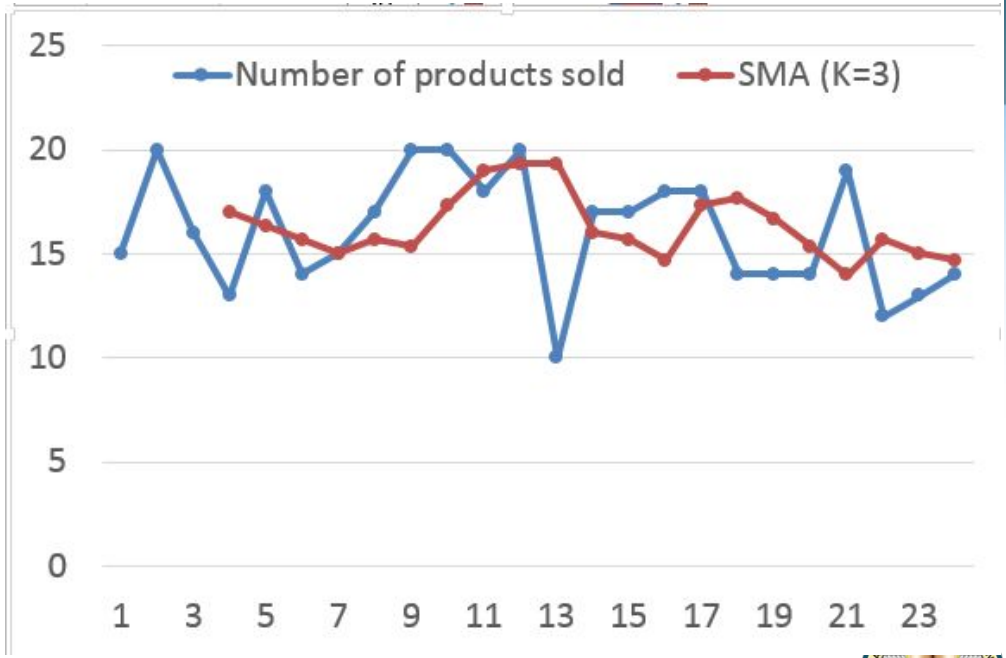
$$\widehat{Y}_{t+1} = \frac{Y_t + Y_{t-1} \cdots + Y_{t-K} + Y_{t-K+1}}{K}$$

- i.e. predicted value at time (t+1) is the average of past K values
- K is the only “tunable parameter”

# Simple Moving Average (SMA) results



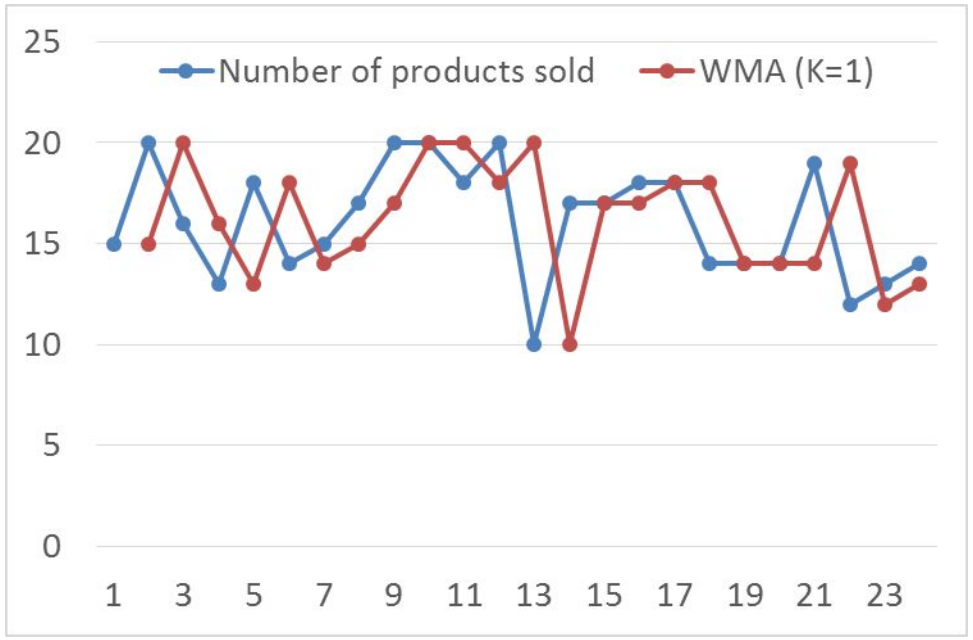



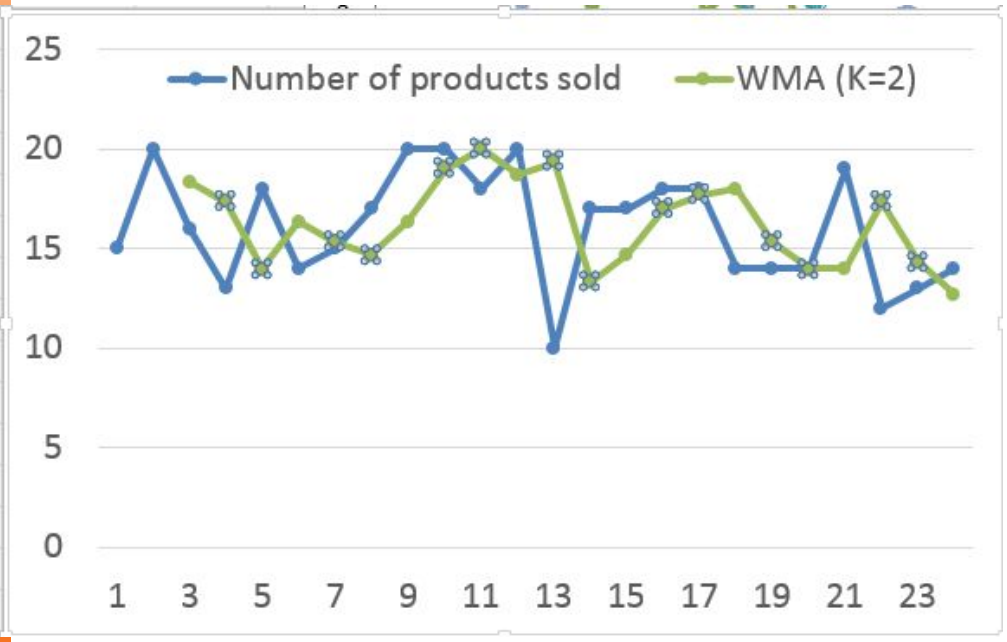
# Stationary Model: Weighted Moving Averages

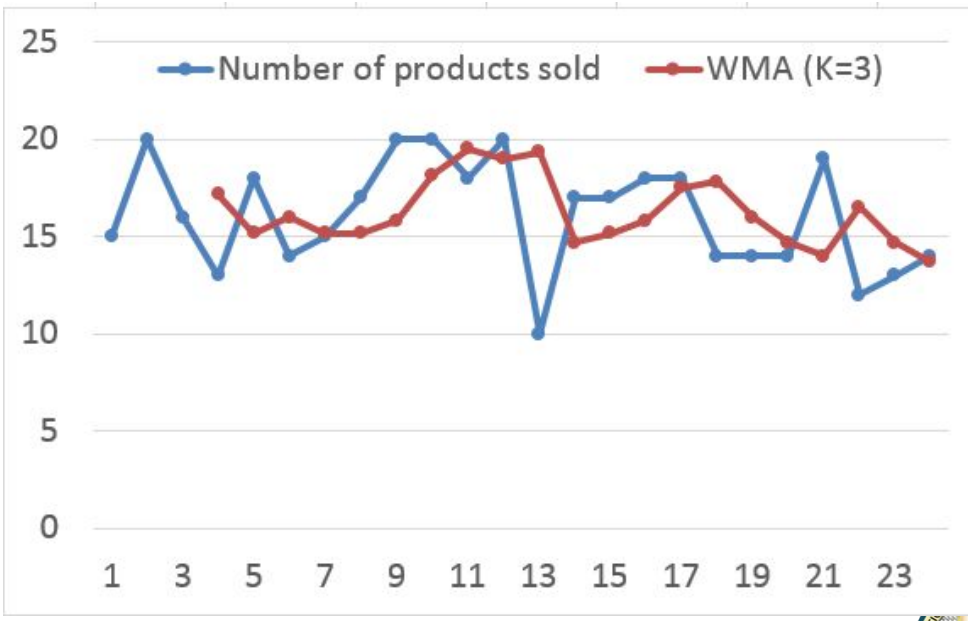
$$\hat{Y}_{t+1} = w_1 Y_t + w_2 Y_{t-1} + \cdots + w_k Y_{t-k+1}$$

- Typically we choose a time period of moving average and weights are chosen such that the error is minimized

# Weighted Moving Average (WMA) results



# Stationary Model: Case 3 – Exponential ~~Weighted Moving Averages~~ or Exponential Smoothing

- Averaging over long periods dampens fluctuations, removing not only the noise but also trend and seasonality.

- Moving averages over short recent periods maintains trend and seasonality but determining an optimum number for periods is tricky, even when using metrics like MAE. If averaged over too few periods, irregularities continue to remain and if averaged over long periods, dampening again becomes a problem.

- Exponential smoothing **retains all older periods** while giving a greater weight to more recent periods (hence not a MOVING average).

- *Caution: It doesn't make any one method superior for all situations.*

# Stationary Model: Case 3–Exponential Smoothing

$$\hat{Y}_{t+1} = \hat{Y}_t + \alpha\left(Y_t - \hat{Y}_t\right)$$

Above equation indicates that the predicted value for time period $t$+1 $(\hat{Y}_{t+1})$ is equal to the predicted value for the previous period $(\hat{Y}_t)$ plus an adjustment for the error made in predicting the previous period's value $(\alpha\left(Y_t - \hat{Y}_t\right))$.

The parameter α can assume any value between 0 and 1 ($0 \leq \alpha \leq 1$).

# Exponential Smoothing in Other Ways

$\hat{Y}_{t+1} = \hat{Y}_t + \alpha(Y_t - \hat{Y}_t)$ can be rewritten variously as

$$= \alpha Y_t + (1-\alpha)\hat{Y}_t$$

$$= Y_t - (1-\alpha)(Y_t - \hat{Y}_t)$$

$$= \alpha Y_t + \alpha(1-\alpha)Y_{t-1} + \alpha(1-\alpha)^2 Y_{t-2} + \cdots + \alpha(1-\alpha)^n Y_{t-n} + \cdots$$

# Exponential Smoothing

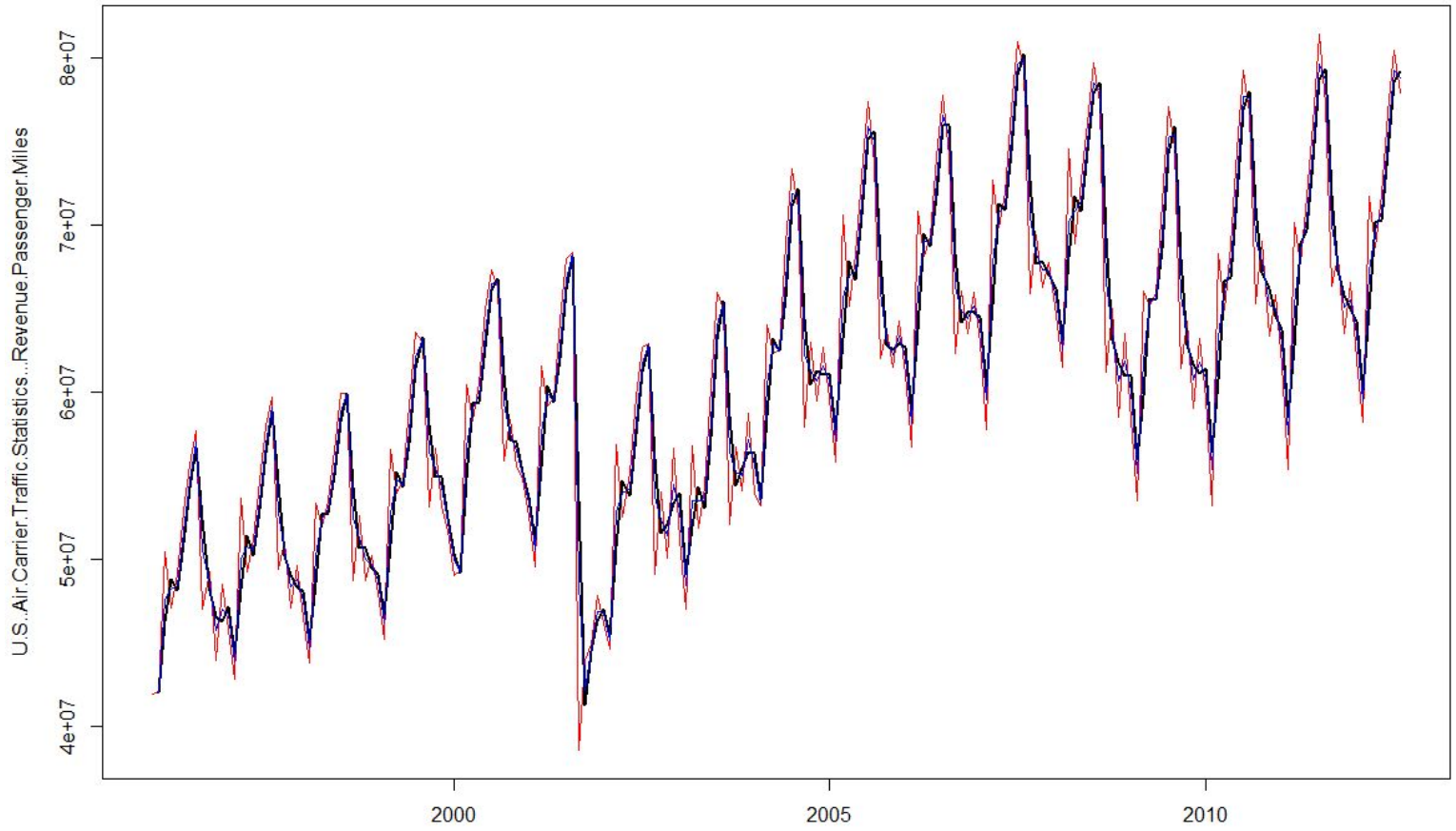
$$\hat{Y}_{t+1} = \hat{Y}_t + \alpha(Y_t - \hat{Y}_t) \qquad \alpha = \frac{2}{N+1}$$

- Forecasting at time t+1 requires both the forecasted value and the True Value at time t

- So if you want to forecast more than 1 time period into the future, the best you can do is to use the last available value

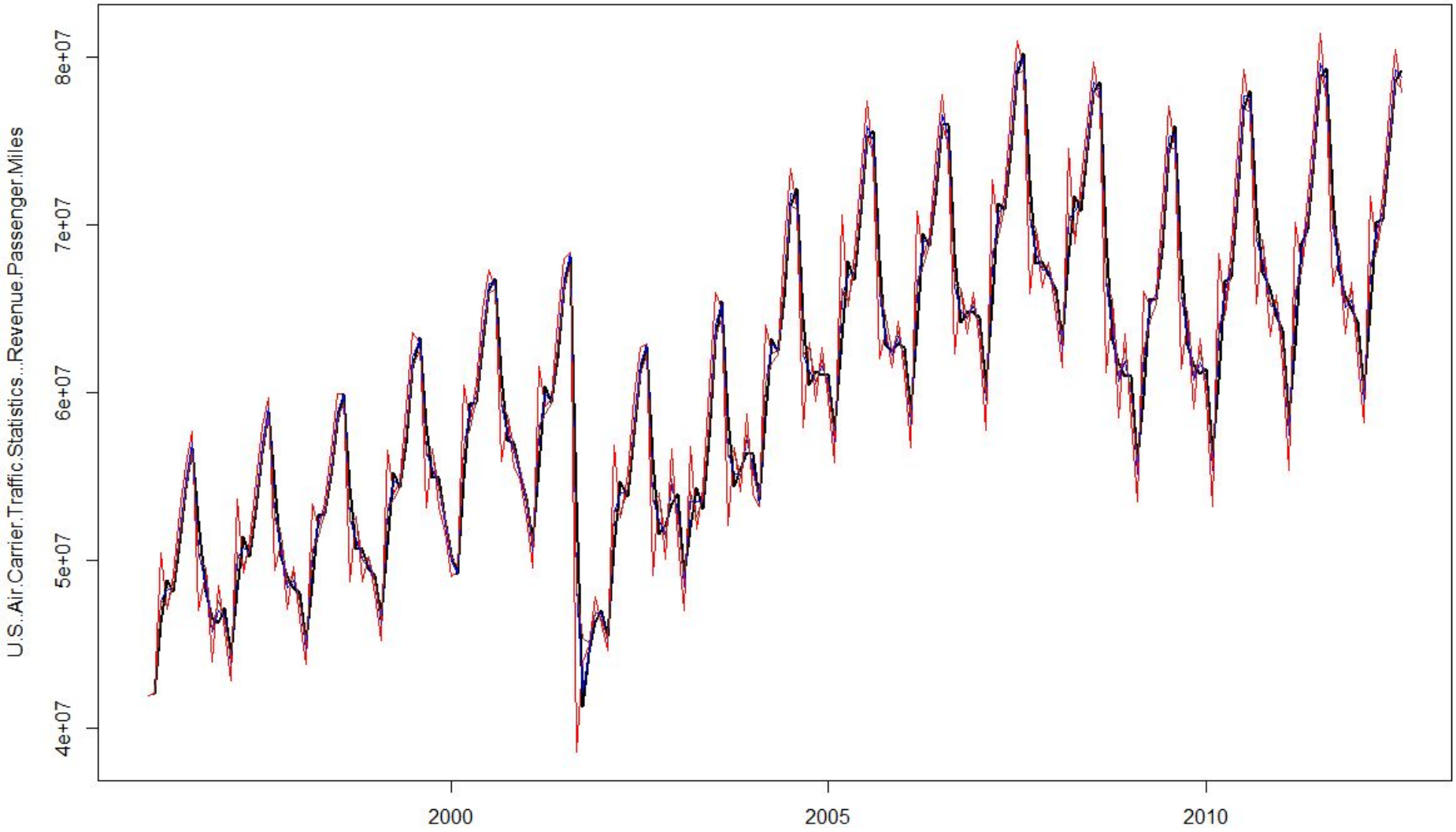
- All future predictions are same! This is in accordance with **stationary** assumption.

# SMA and WMA – Revenue Passenger Miles




> MAPE-SMA 4.093731 > MAPE-WMA 2.729154

# SMA, WMA and Exponential Smoothing – RPM




> MAPE SMA 4.093731 > MAPE WMA 2.729154 > MAPE EMA 2.541979

# AR, MA AND ARIMA MODELS

# AR(p) models

- Auto-regressive model of order p

$$\hat{y}_t = \alpha + \beta_1 y_{t-1} + \beta_2 y_{t-2} + \cdots + \beta_p y_{t-p}$$

- We find the best value of parameters ($\beta_1$, $\beta_2$,…) that minimize the errors in forecast of $\hat{y}_t$.

- The order of the model p is determined based on the number beyond which PACF terms are zero.

# Moving Average or MA(q) models

- Model attempts to predict future values using past error in predictions. Eg. $\varepsilon_1 = \hat{y}_1 - y_1$
- For example, MA(2) model is $\hat{y}_t = \mu + \phi_1 \varepsilon_{t-1} + \phi_2 \varepsilon_{t-2}$
  - where $\mu$ is the average value of the time series.
  - The parameters ($\phi_1$,$\phi_2$) are determined so that prediction error is minimized.
- The number of terms, q, is determined from the ACF plot. It is the maximum lag beyond which the ACF is 0

# Use of ACF and PACF to identify AR and MA models

**Identification of an AR model often best done with the PACF.**

- For an AR model, theoretically the number of non-zero partial autocorrelations gives the order of the AR model.
- Here "order of the model" refers to the most extreme lag of time series variable that is used as a predictor.

**Identification of an MA model is often best done with the ACF**
(A clearer pattern for an MA model is in the ACF rather than in PACF)

- The ACF will have non-zero autocorrelations only at lags involved in the model.

# ARMA(p,q) model

- Combines both AR(p) and MA(q) models
- So a ARMA(2,1) model is

$$\hat{y}_t = \alpha + \beta_1 y_{t-1} + \beta_2 y_{t-2} + \phi_1 \varepsilon_{t-1}$$

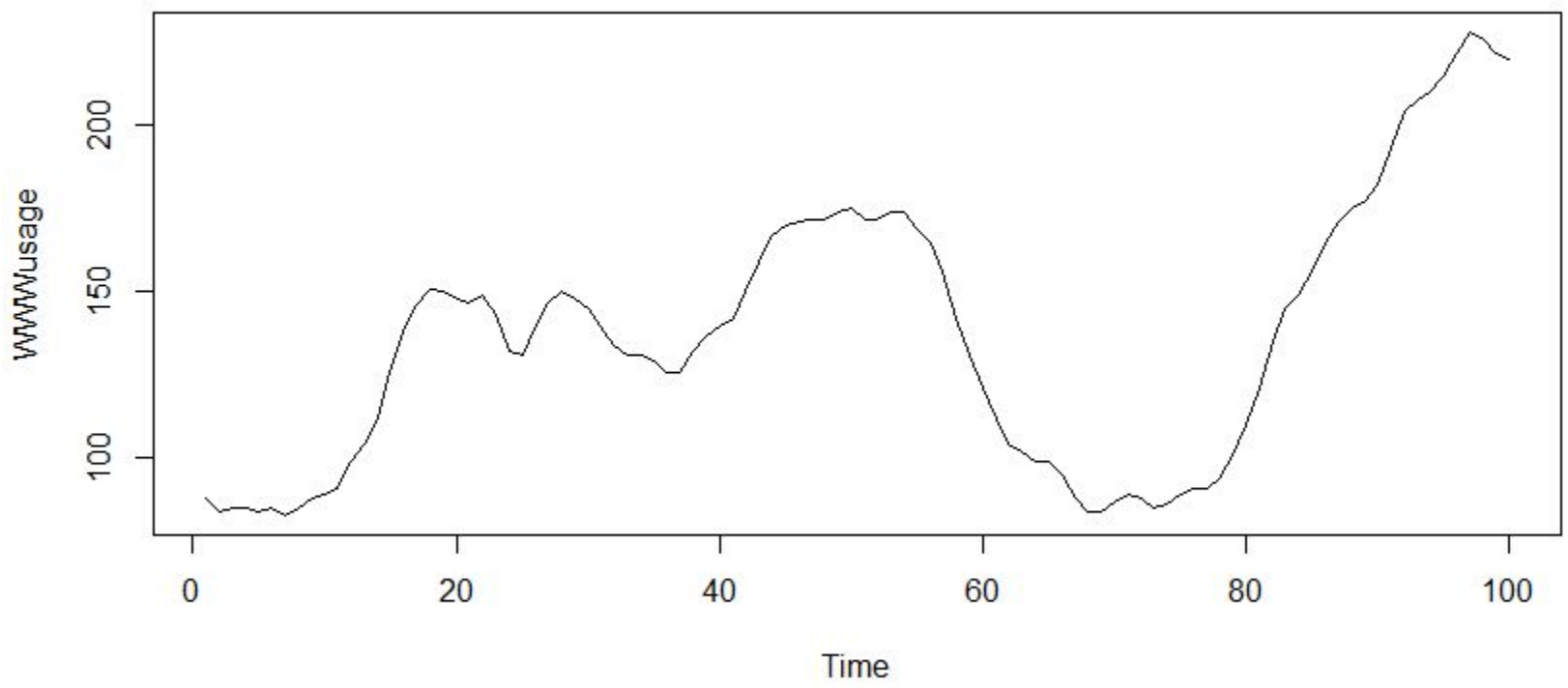
AR (p=2) terms

MA (q=1) term

# ARIMA(p,d,q) Model

- *p* is the number of autoregressive terms (a linear regression of the current value of the series against one or more prior values of the series)
  - Maximum lag beyond which PACF is 0
- *d* is the number of non-seasonal differences (order of the differencing) used to make the time series stationary
- *q* is the number of past prediction error terms used for the future forecasts.

# Using ARIMA to forecast

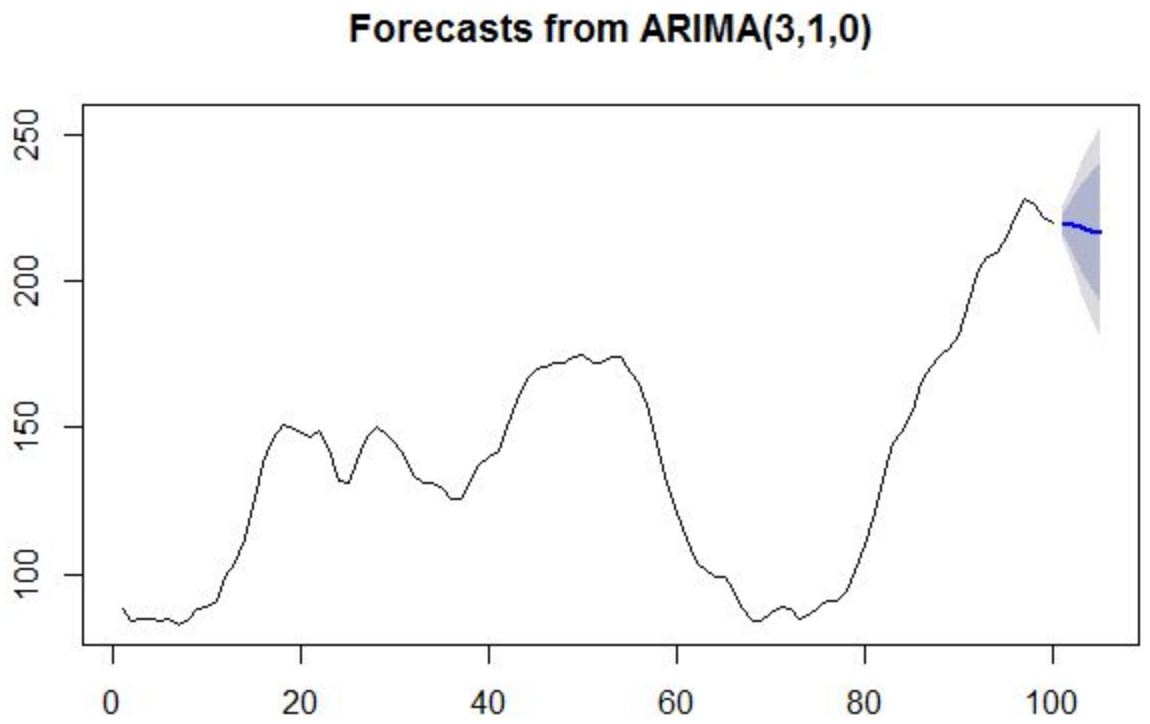



A time series of the numbers of users connected to the Internet through a server every minute.

# Using ARIMA to forecast

Let us see what the forecast looks like if we use Arima(3,1,0) model

```
>
> fit <- Arima(WWWusage,c(3,1,0))
> plot(forecast(fit,h=5))
>
```




The forecast is plotted in dark blue.
The dark grey and light grey regions representthe 80% and 95% confidence intervals.

# Model Identification

- Before Automated functions were available, one needed to use ACF plots to determine the best value of (p,d,q) for a given dataset

- Box–Jenkins Methodology: Model identification and model selection(http://www.ncss.com/wp-content/themes/ncss/pdf/Procedures/NCSS/The_Box-Jenkins_Method.pdf)
  - Make sure variables are stationary.
    - **A diagnostic tool for stationarity : Dickey-Fuller tests**
  - Difference as necessary to get a constant mean and transformations to get constant variance.
  - Check for seasonality: Decays and spikes at regular intervals in ACF plots.

# Model Identification

- Parameter estimation
  - Compute coefficients that best fit the selected model.

- Model checking
  - Check if residuals are independent of each other and constant in mean and variance over time (white noise).
  - Testing for white noise
    - Box-Pierce statistic, Ljung-Box statistic

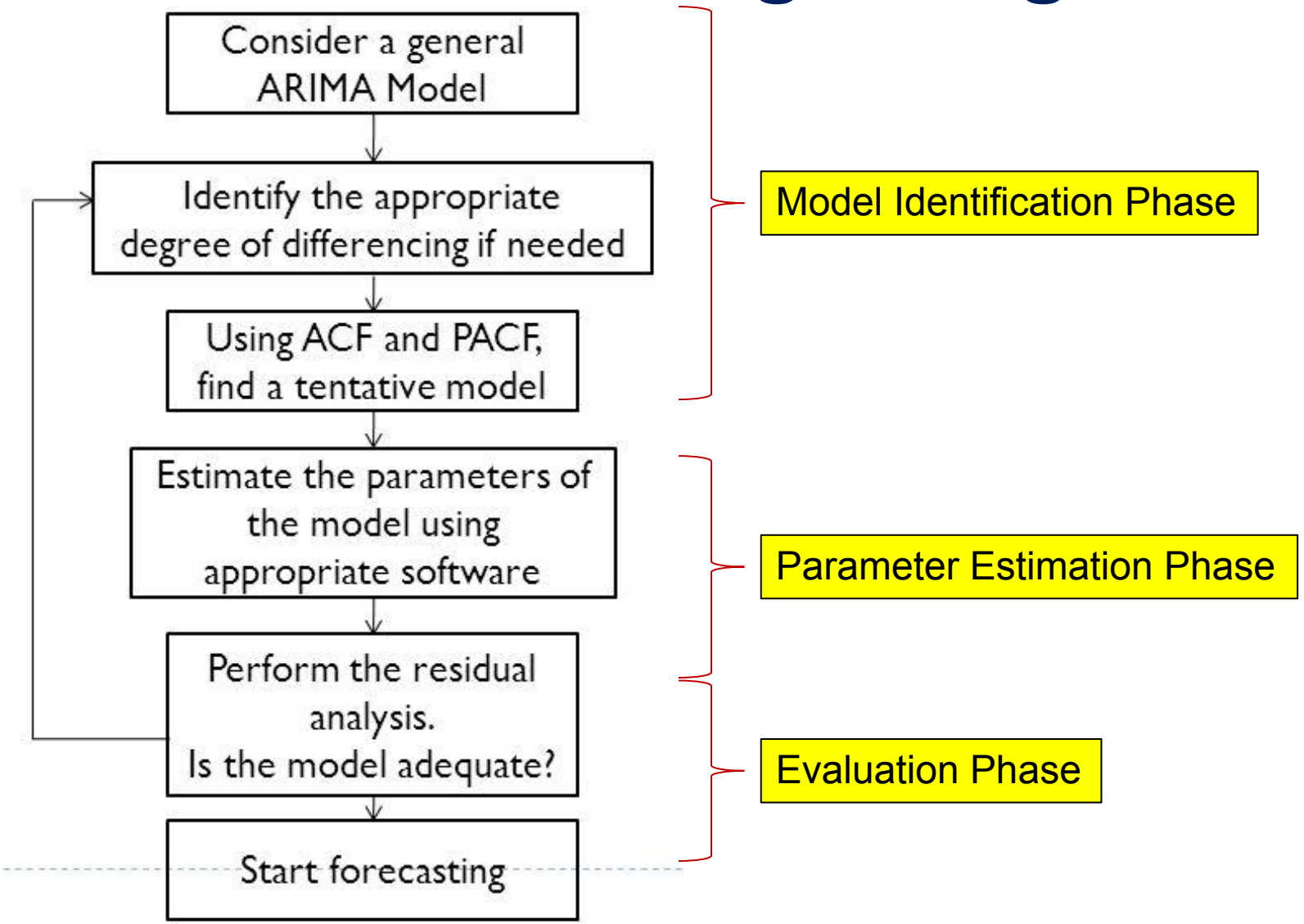
- Goodness of fit
  - Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) are two commonly used goodness of fit measures.

# ARIMA and SARIMA

- Non-seasonal ARIMA models are denoted ARIMA(p,d,q)

- Seasonal ARIMA (SARIMA) models are denoted ARIMA(p,d,q)$(P,D,Q)_m$, where m refers to the number of periods in each season and (P,D,Q) refer to the autoregressive, differencing and moving average terms of the seasonal part of the ARIMA model.

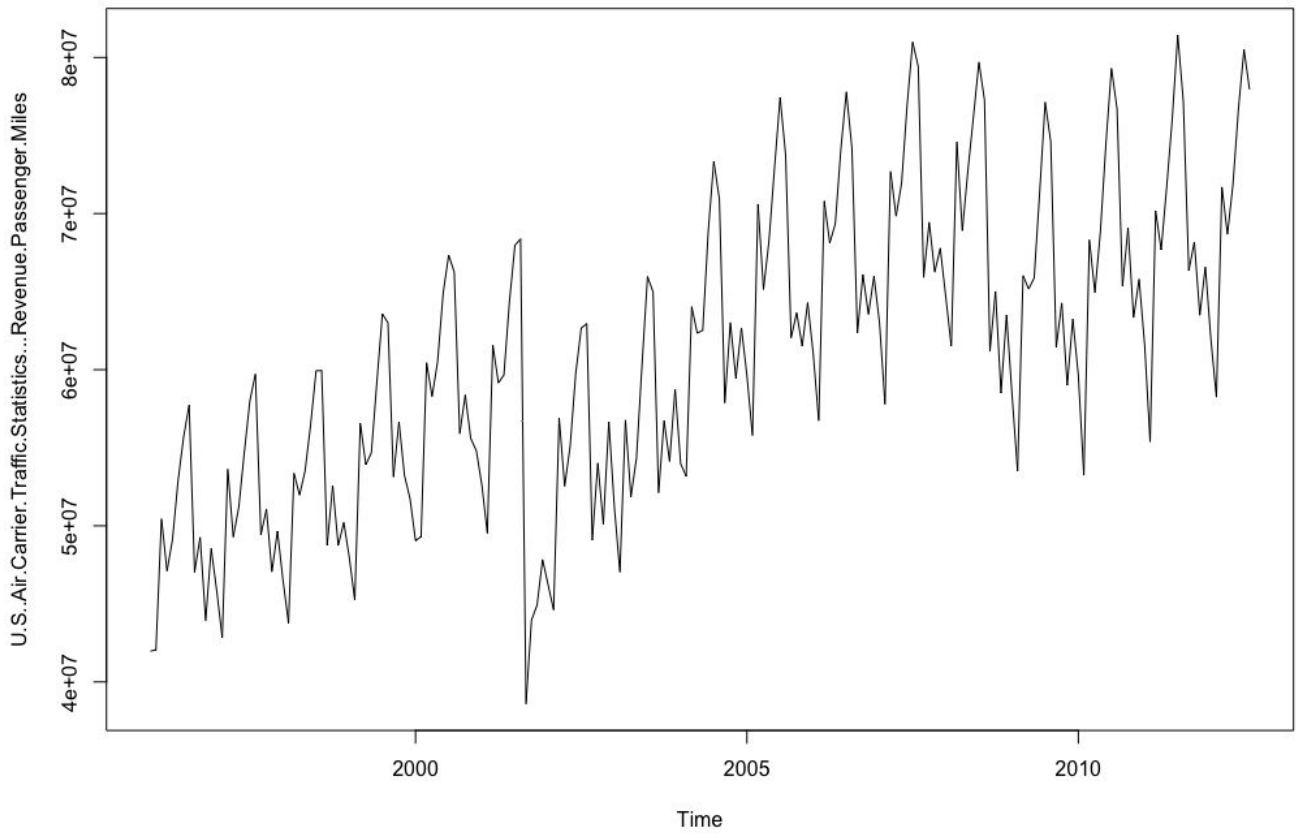
# Time Series Model Building Using ARIMA



# Time Series Model Building with ARIMA

## Identification Phase

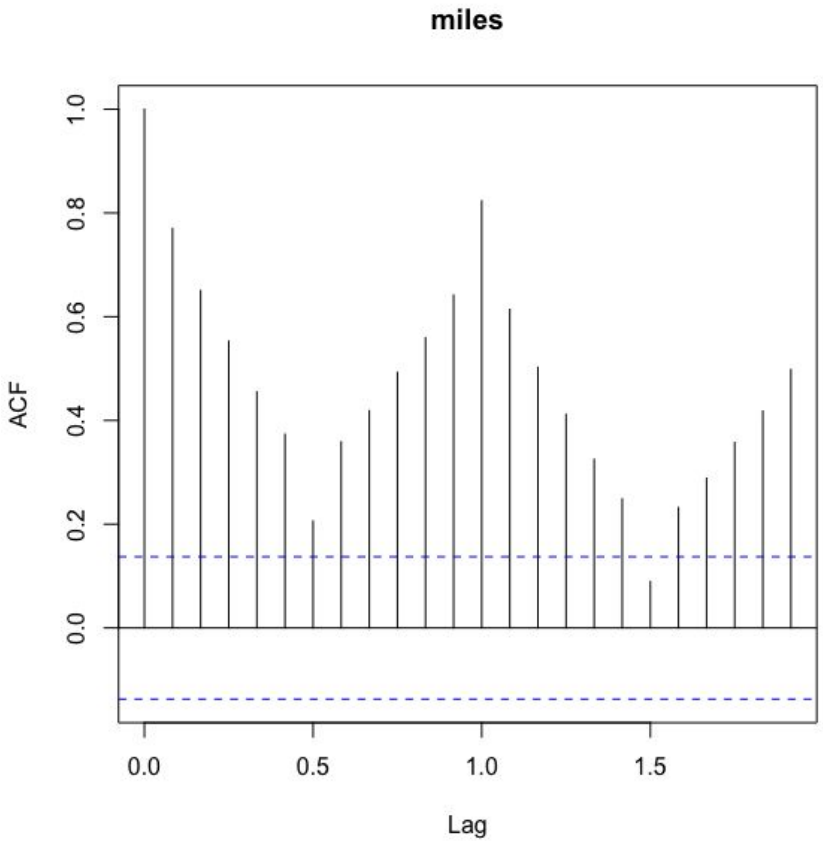
Step 1: Plot the data (transform data to stabilize variance, if required)

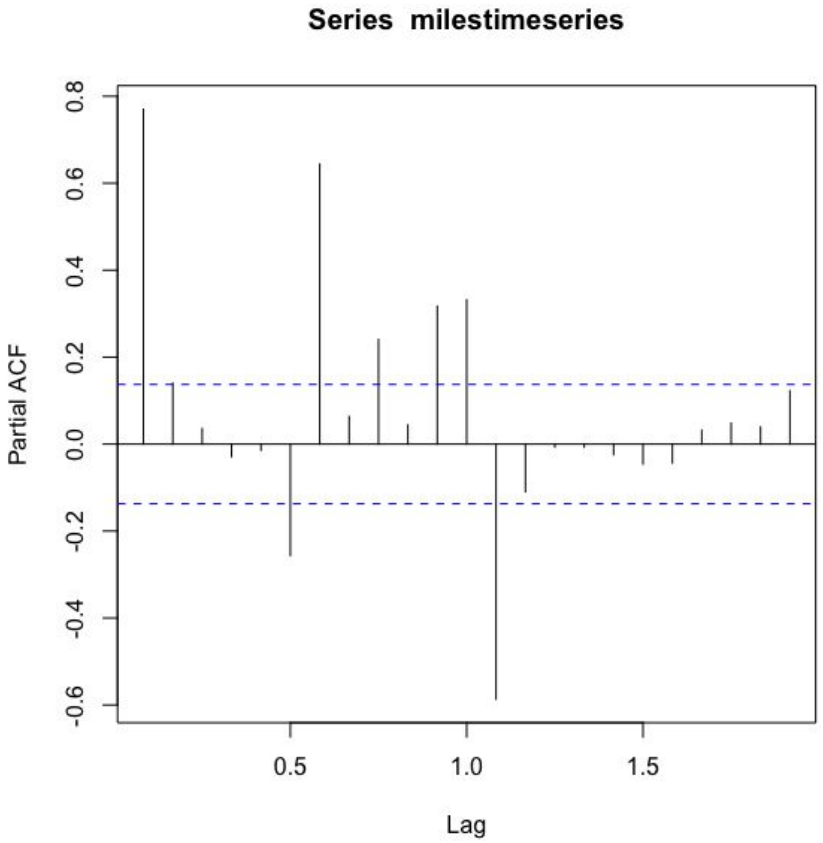


# Time Series Model Building Using ARIMA

## Identification Phase

Step 2: Plot ACF and PACF to get preliminary understanding of the processes involved.

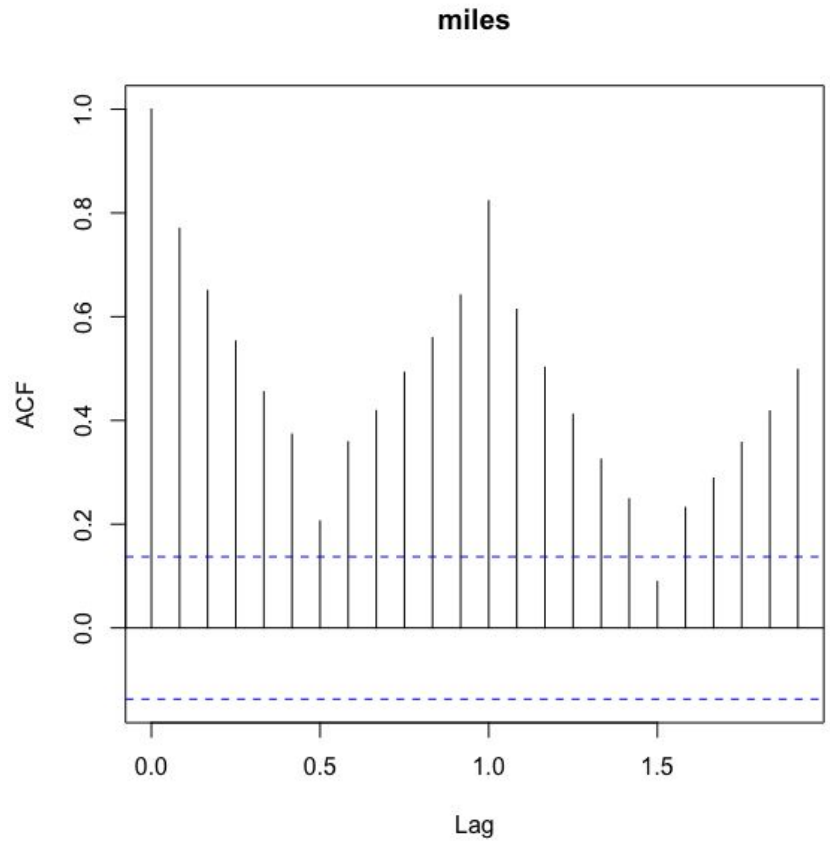


# Time Series Model Building Using ARIMA

## Identification Phase

Step 2: Plot ACF and PACF to get preliminary understanding of the processes involved.

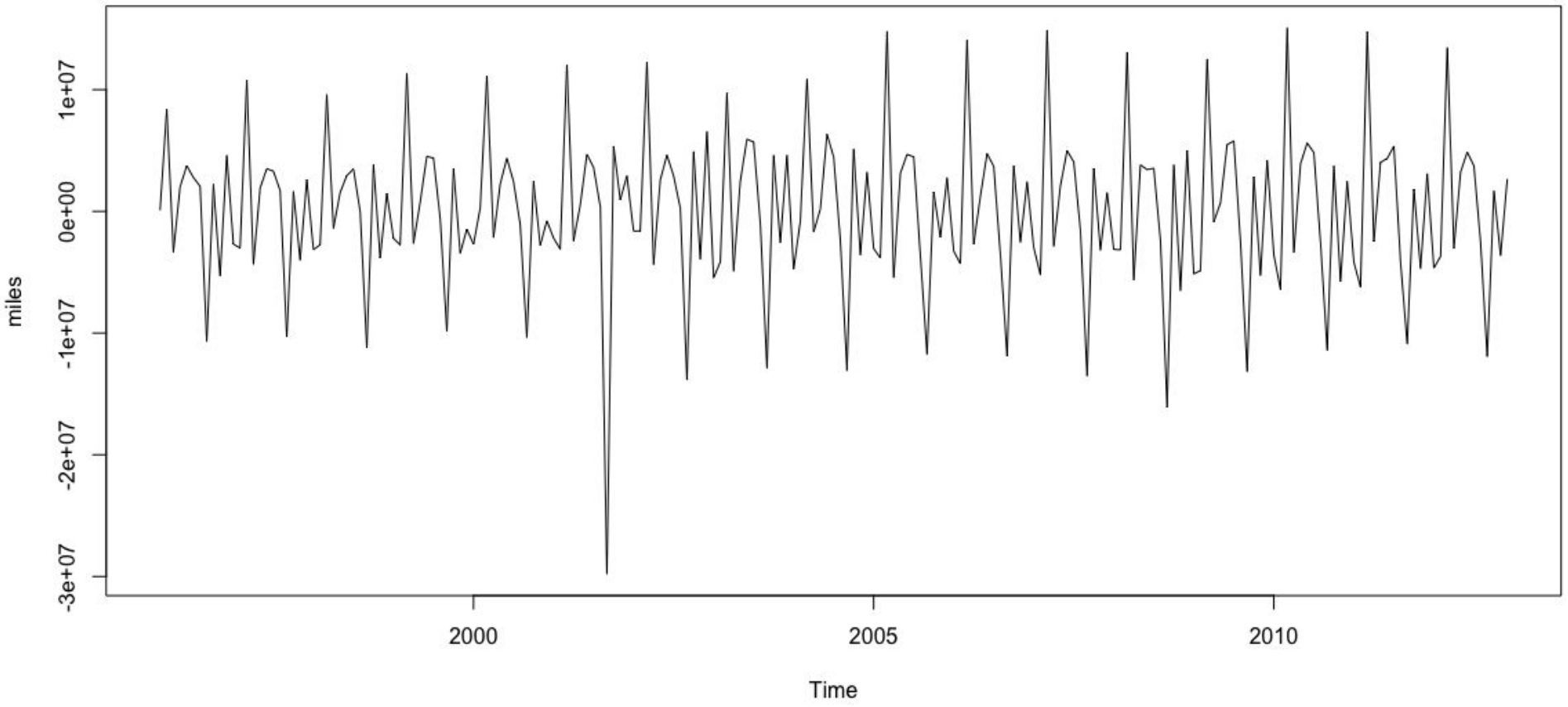



The suspension bridge pattern in ACF (also, positive and negative spikes in PACF) suggests non-stationarity and strong seasonality.

# Time Series Model Building Using ARIMA

## Identification Phase

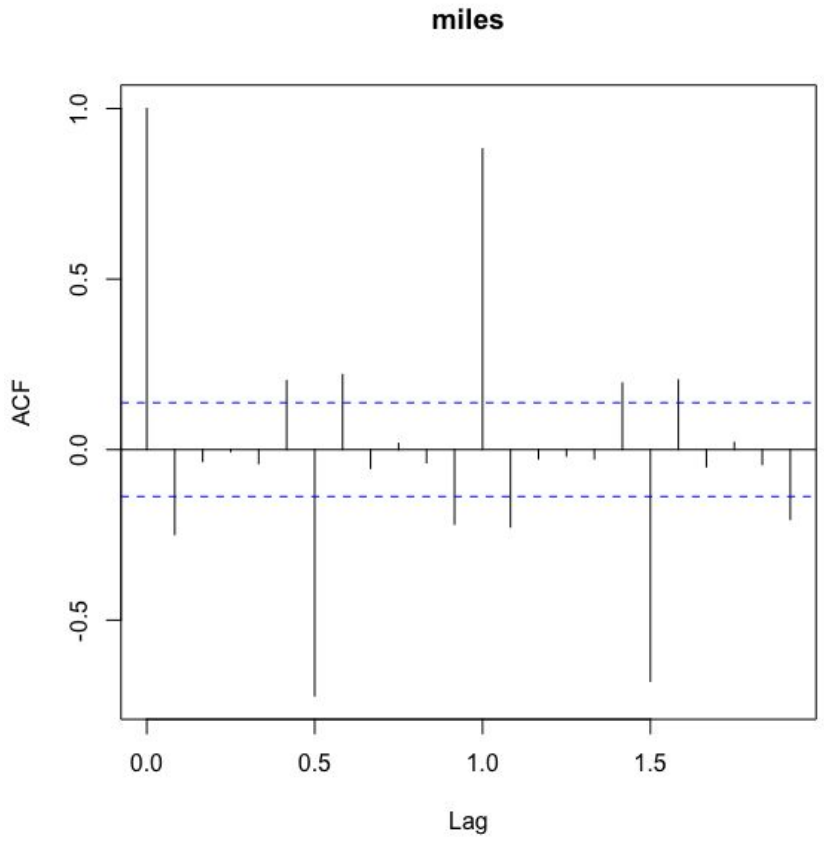
Step 3: Perform a non-seasonal difference. It is the same as an ARIMA(0,1,0) model.

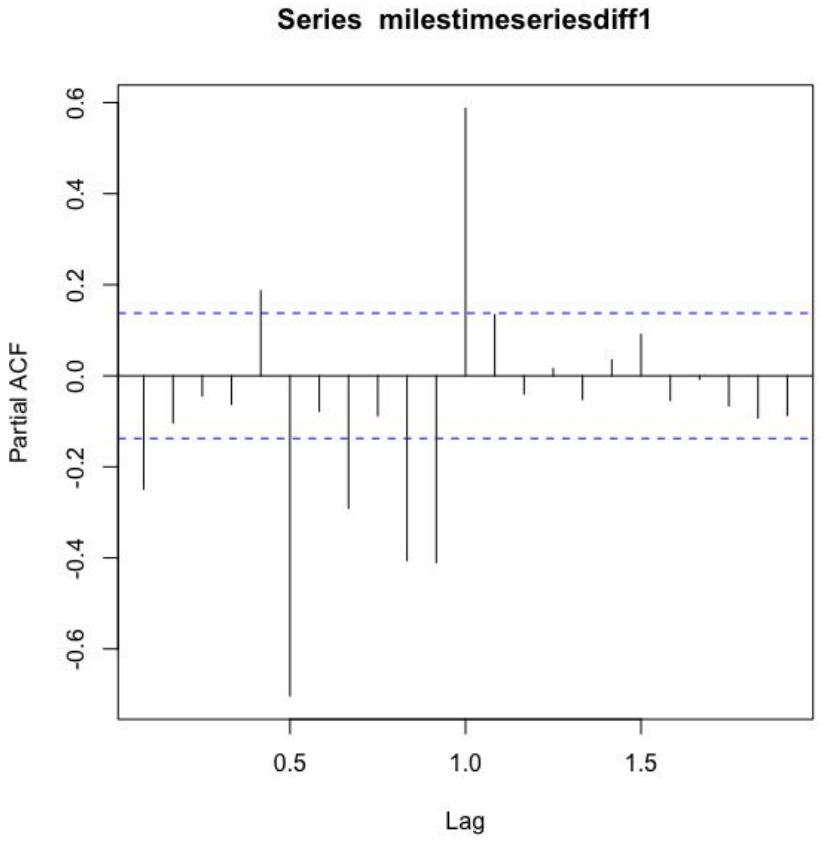


# Time Series Model Building Using ARIMA

## Identification Phase

Step 4: Check ACF and PACF of differenced data to explore remaining dependencies.



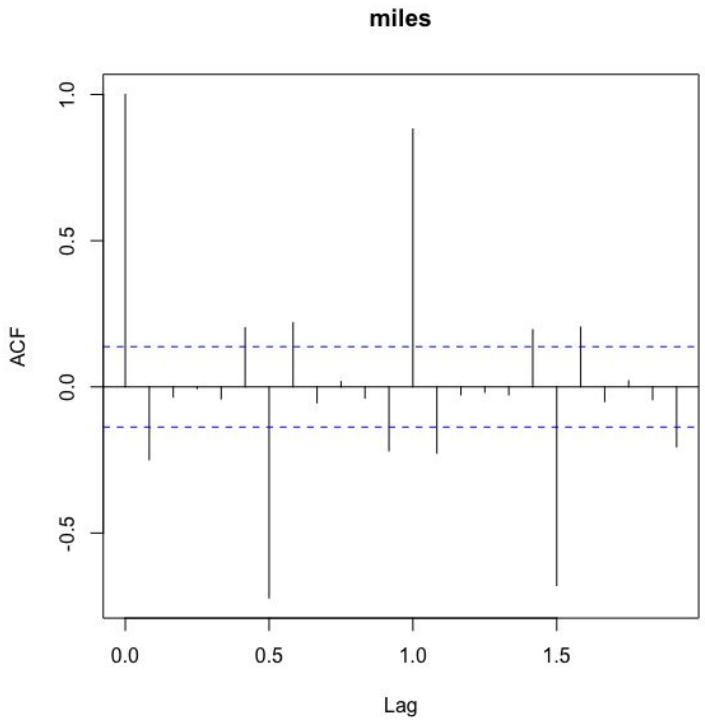



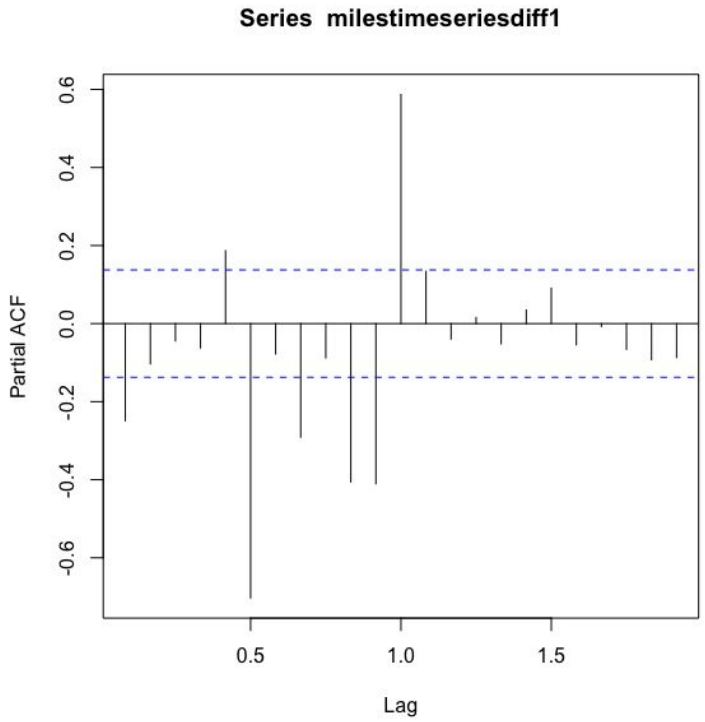
# Time Series Model Building Using ARIMA

## Identification Phase

Step 4: Check ACF and PACF of differenced data to explore remaining dependencies.

The differenced series looks somewhat stationary but has strong seasonal lags.

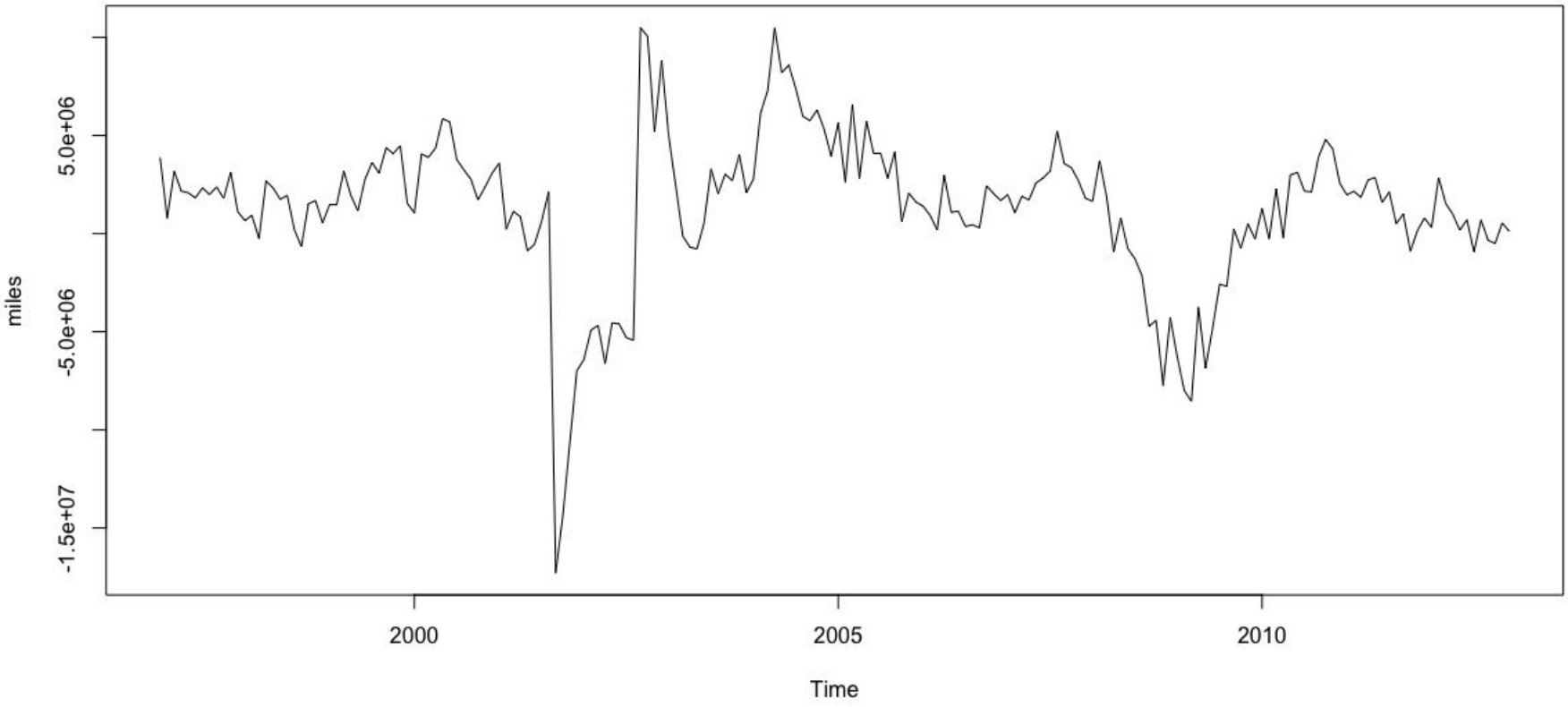


# Time Series Model Building Using ARIMA

## Identification Phase

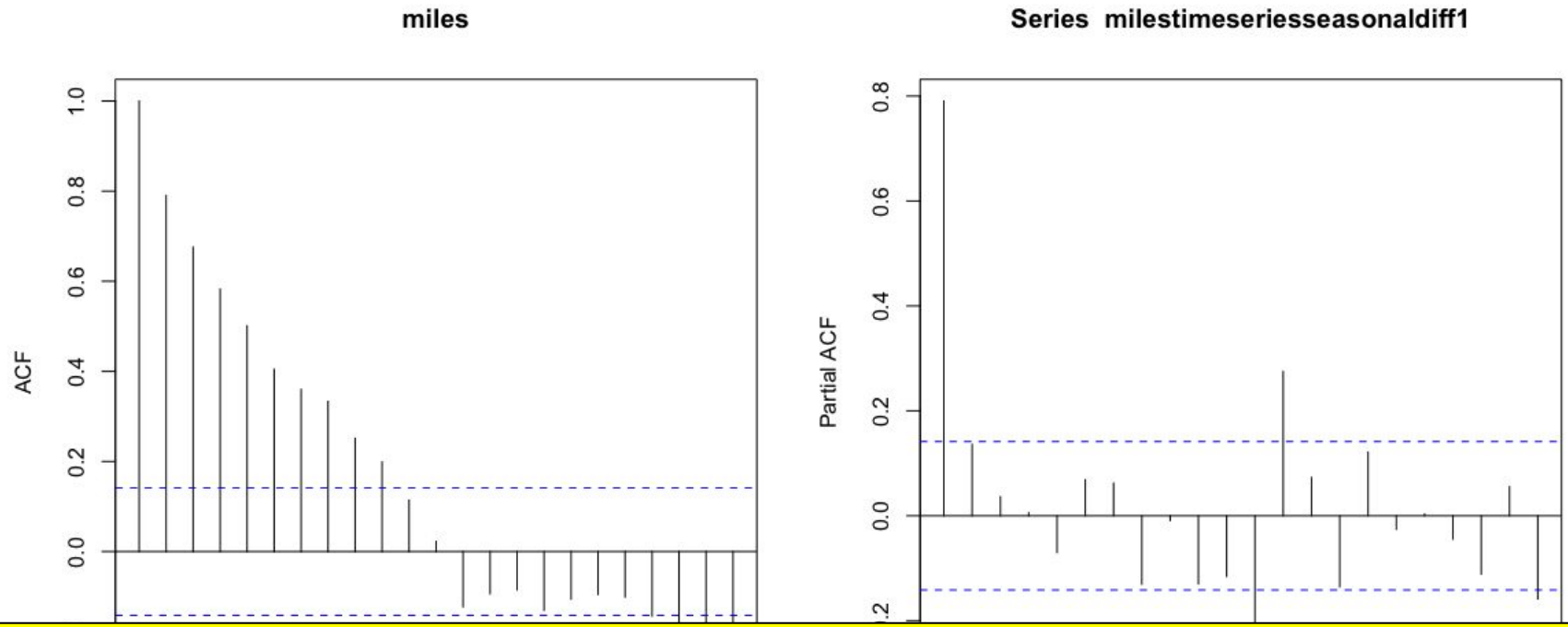
Step 5: Perform seasonal differencing ($t_0$-$t_{12}$, $t_1$-$t_{13}$, etc.) on the **original** time series to get seasonal stationarity. This is the same as an ARIMA(0,0,0)(0,1,0)$_{12}$ model.



# Time Series Model Building Using ARIMA

## Identification Phase

Step 6: Check ACF and PACF of seasonally differenced data to explore remaining dependencies and identify model(s).

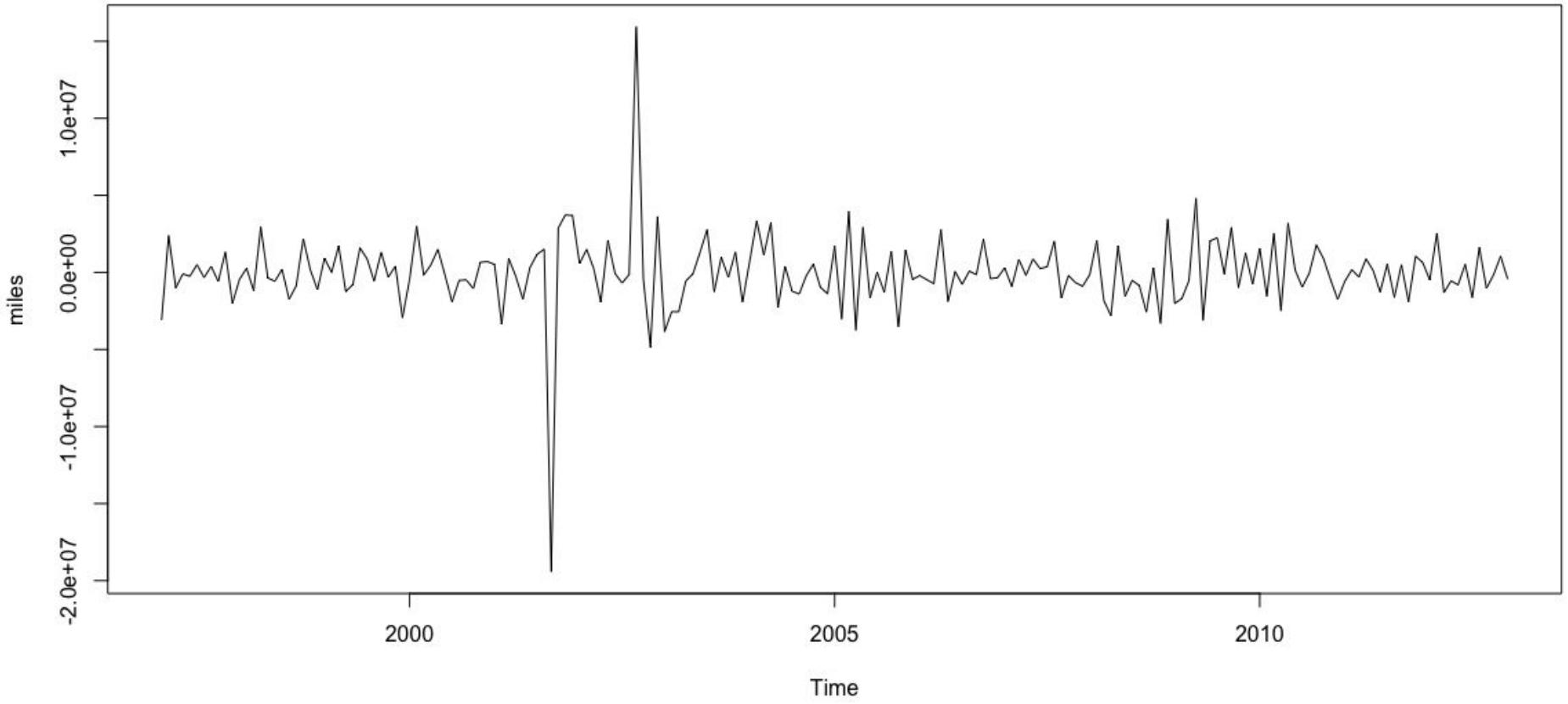



Strong positive autocorrelation indicates need for either an AR term or a non-seasonal differencing.

# Time Series Model Building Using ARIMA

## Identification Phase

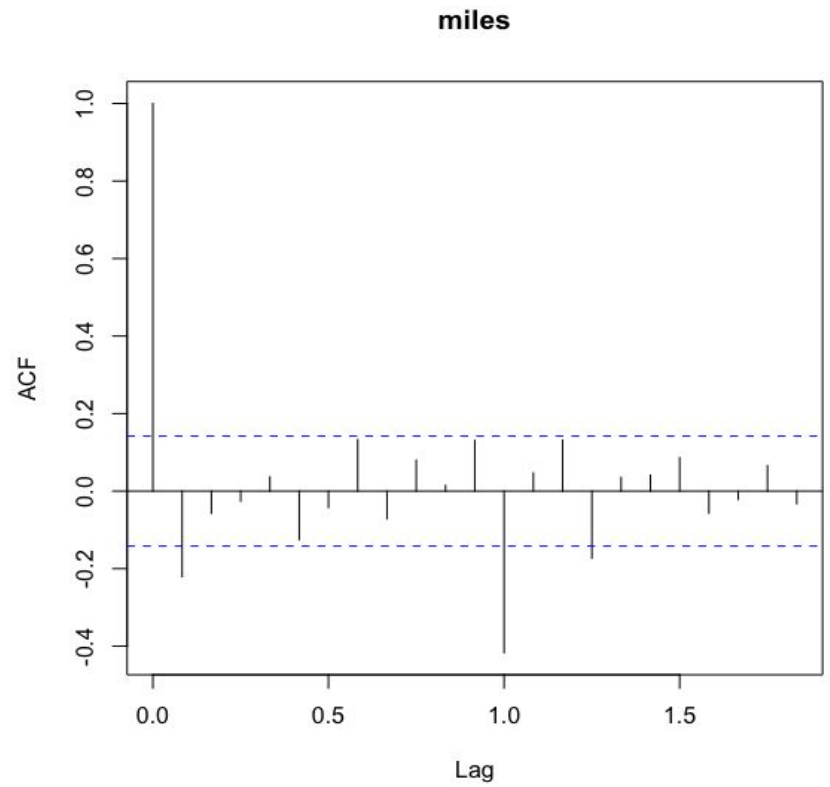
Step 7: Perform a non-seasonal differencing on **seasonally differenced data**. This is like an ARIMA(0,1,0)(0,1,0)$_{12}$ model.

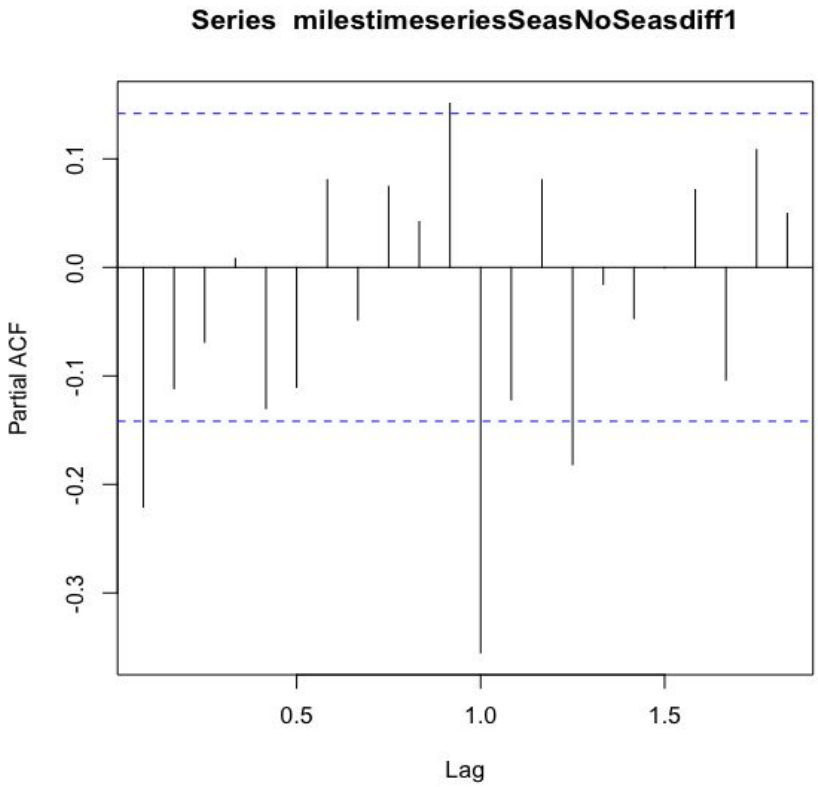


# Time Series Model Building Using ARIMA

## Identification Phase

Step 8: Check ACF and PACF to explore remaining dependencies.

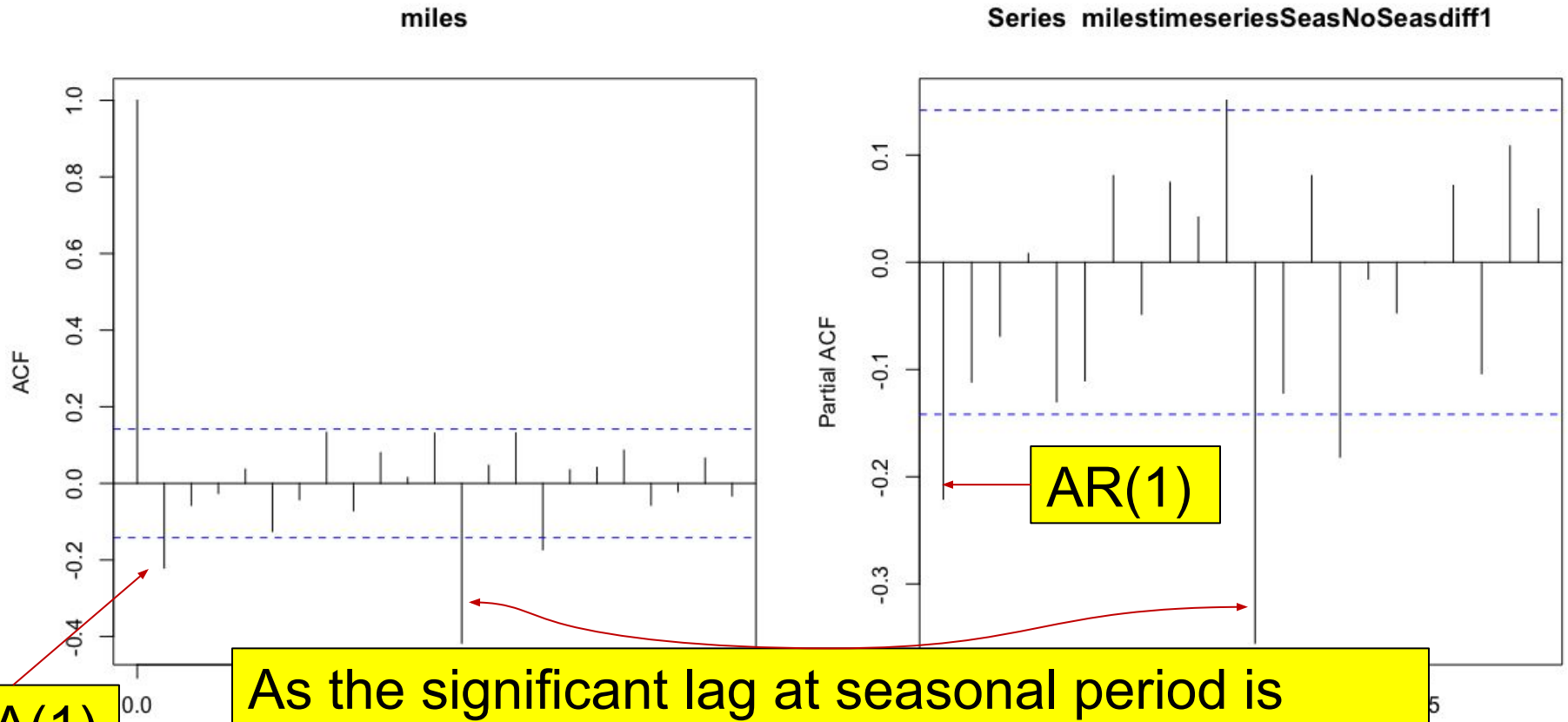


# Time Series Model Building Using ARIMA

## Identification Phase

Step 8: This indicates an ARIMA(1,1,1)(0,1,1)$_{12}$ model.



# Time Series Model Building Using ARIMA

## Parameter Estimation Phase

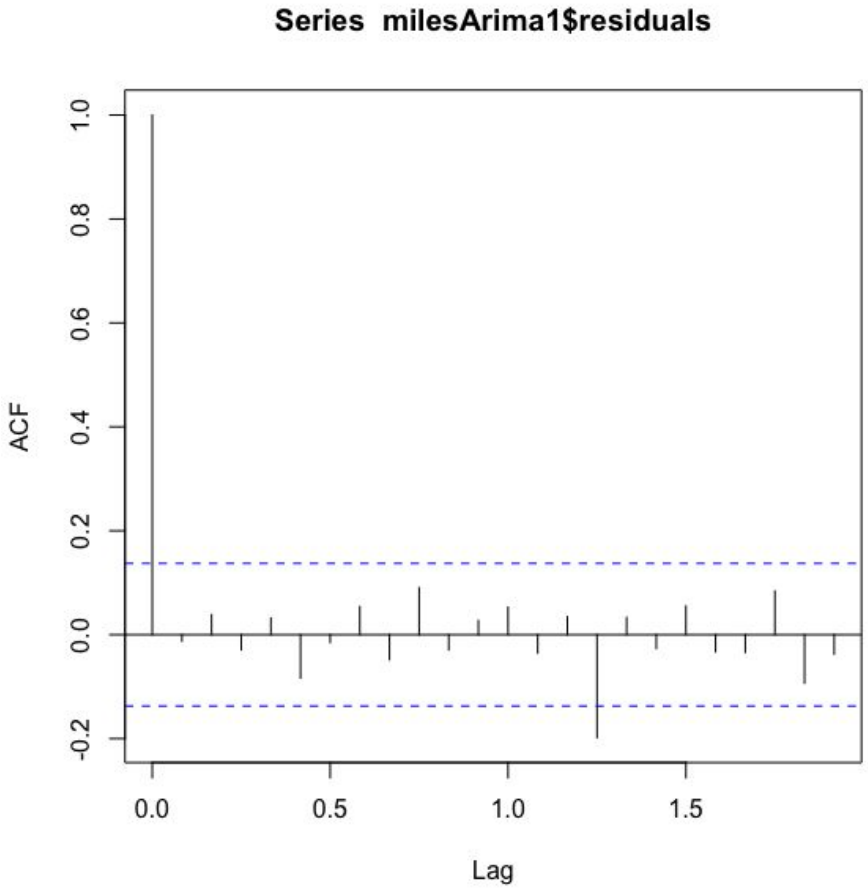
Step 9: Calculate parameters using the identified model(s).
Use AIC to pick the best model.

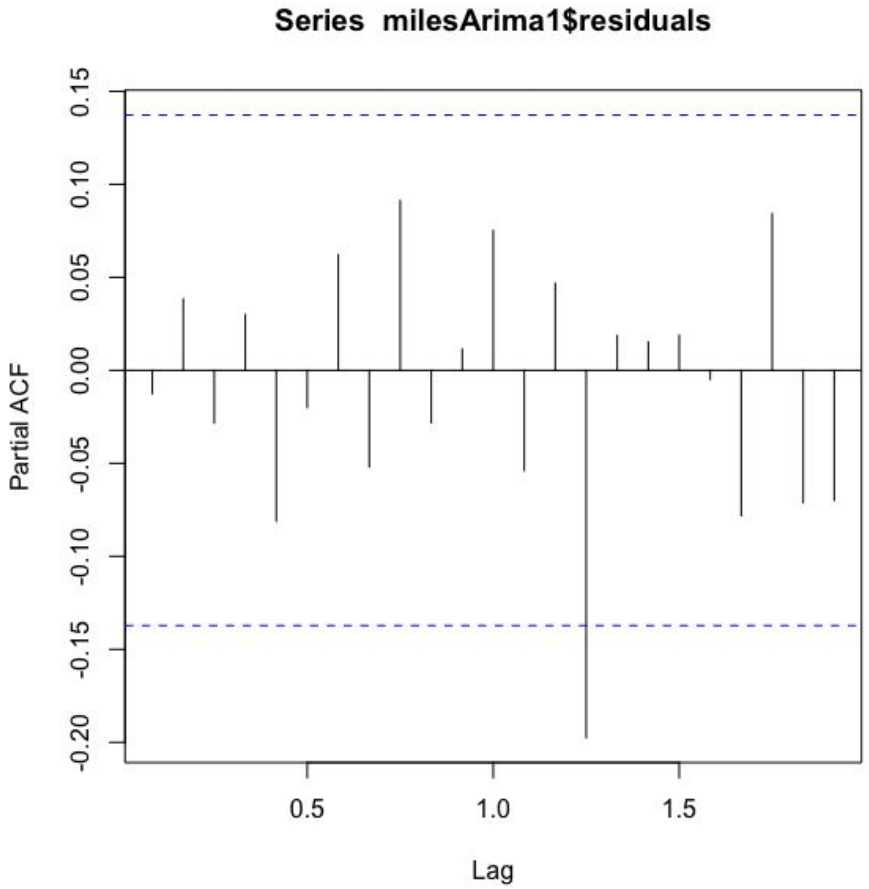
```
Series: milestimeseries
ARIMA(1,1,1)(0,1,1)[12]

Coefficients:
          ar1      ma1     sma1
      0.4501  -0.7035  -0.7393
s.e.  0.1755   0.1407   0.0641

sigma^2 estimated as 3.917e+12:  log likelihood=-3043.49
AIC=6094.99   AICc=6095.2   BIC=6107.99
```

# Time Series Model Building Using ARIMA

## Evaluation Phase

Step 10: Check ACF and PACF of the residuals to evaluate

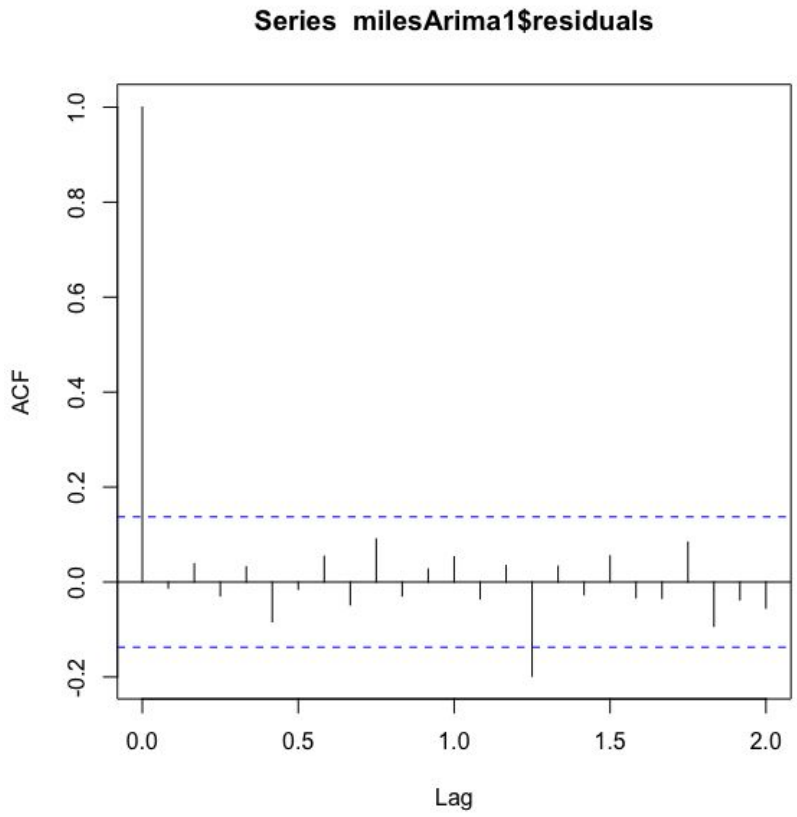


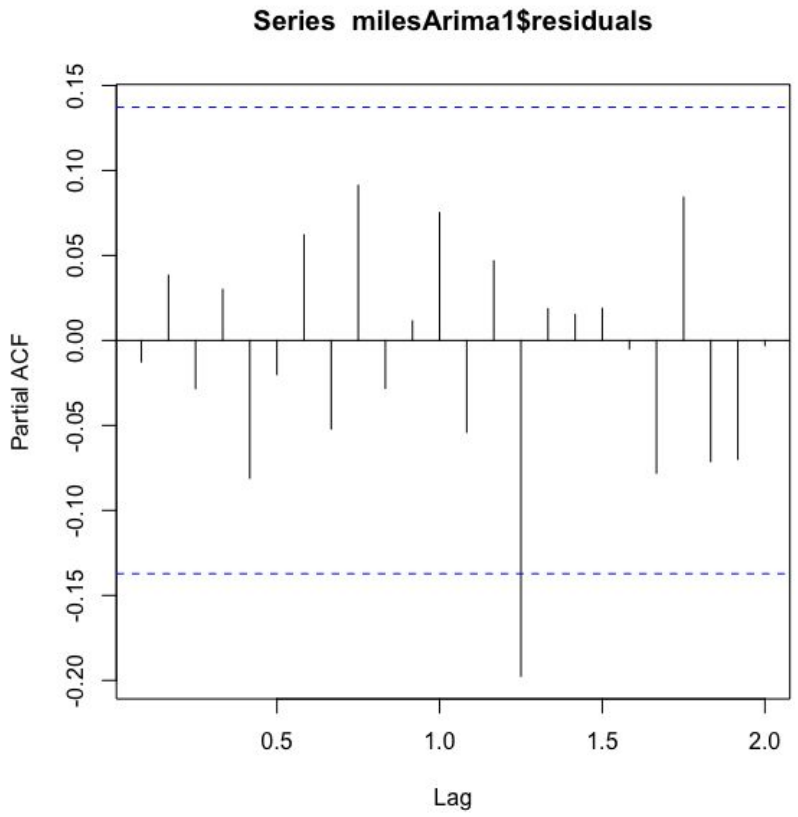



# Time Series Model Building Using ARIMA

## Evaluation Phase

Step 10: The residuals indicate white noise. Indicates a good model that can be used for forecasting.



# Time Series Model Building Using ARIMA

## Evaluation Phase

Step 10: The residuals should indicate white noise if the model fit is good. White noise assumption can be checked by performing Ljung-Box test on the residuals

**$H_0$:** The data are independently distributed (i.e. the correlations in the population from which the sample is taken are 0, any observed correlations in the data result from randomness of the sampling process).
**$H_a$:** The data are not independently distributed; they exhibit serial correlation.

```
        Box-Ljung test

data:  milesArima1$residuals
X-squared = 21.65, df = 24, p-value = 0.6002
```

Result of the Ljung-Box test on the residuals after fitting an AR(1,1,1)(0,1,1) model

# Time Series Model Building Using ARIMA

## Evaluation Phase

Step 10: Ljung-Box test details

$$Q^* = n(n+2)\sum_{k=1}^{h}\frac{{r_k}^2}{n-k}$$

*h* is the maximum lag being considered
*n* is the # of observations (length of the time series)
$r_k$ is the autocorrelation

If residuals are white noise (purely random), then *Q* has a $\chi^2$ distribution with *h-p* degrees of freedom, where *p* is the number of parameters estimated in the model

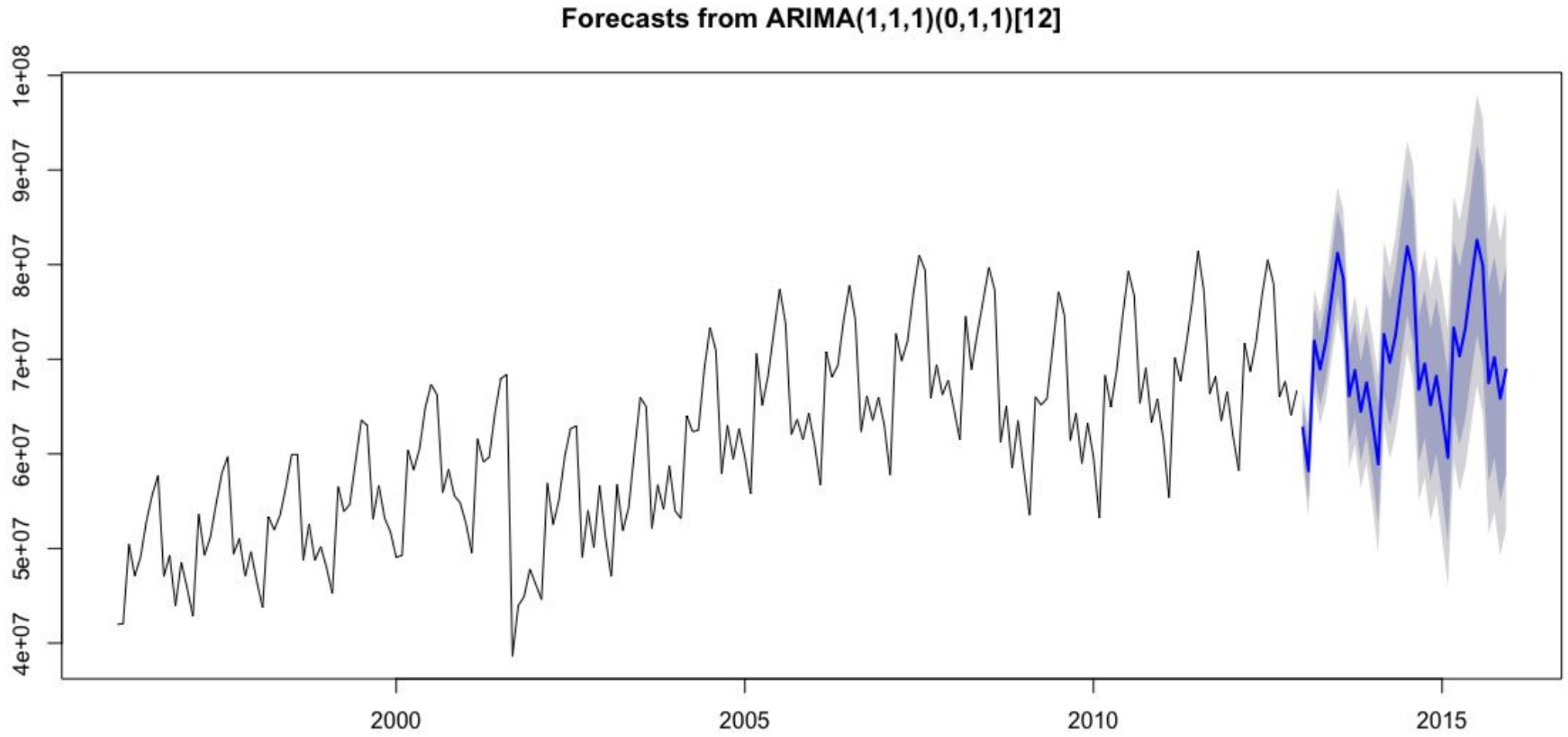
* For non-seasonal time series, use *h = min(10, n/5)*
For seasonal time series, use *h = min(2m, n/5)*, where m is the seasonal period

* http://robjhyndman.com/hyndsight/ljung-box-test/

# Time Series Model Building Using ARIMA

## Forecasting Phase

Step 11: Start forecasting.



# Time Series Model Building Using ARIMA

A nice summary of rules for identifying ARIMA models

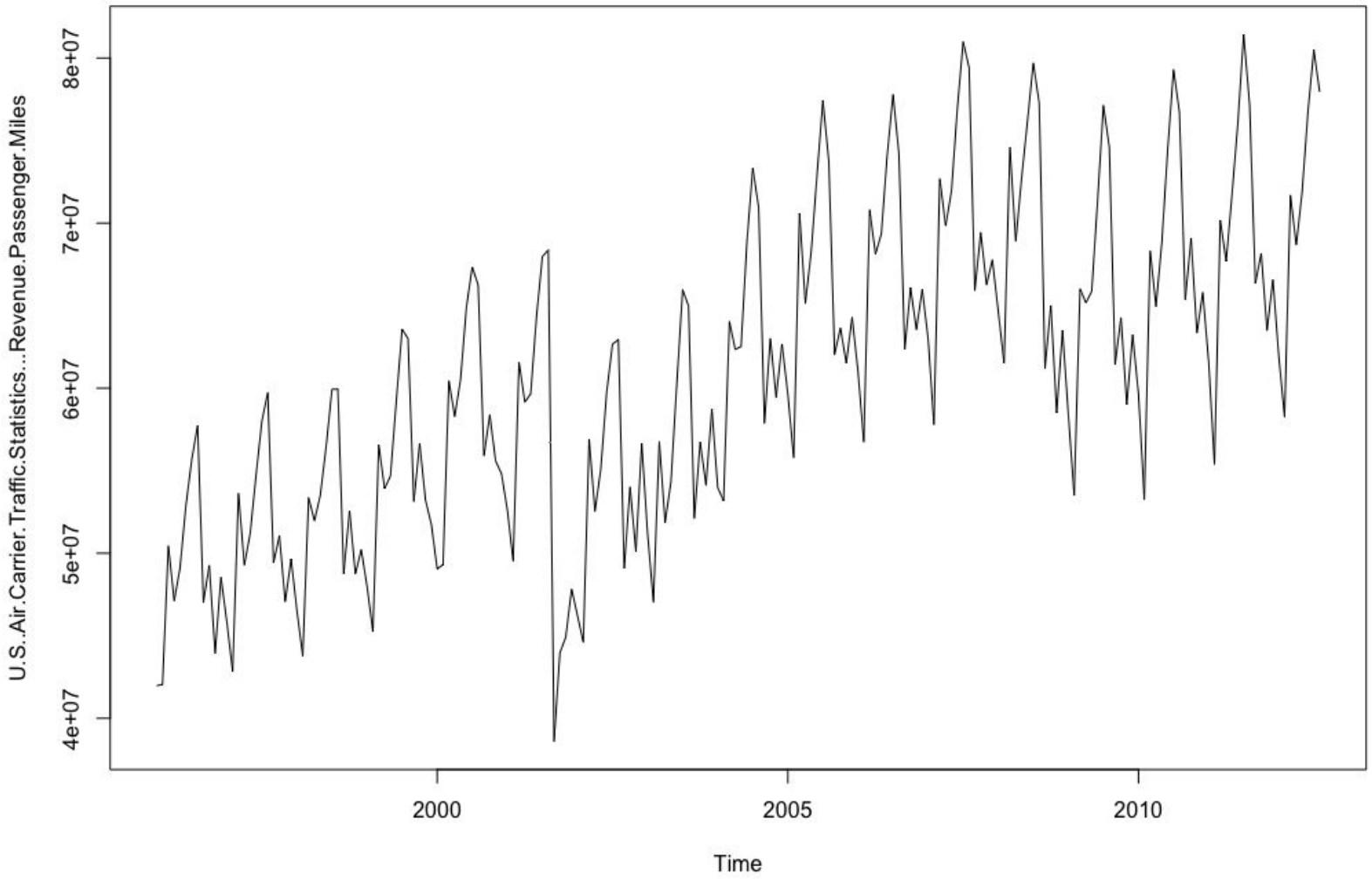
http://people.duke.edu/~rnau/arimrule.htm

# Model Selection

- The number of parameters (p,d,q) needed to fit, depends on the dataset
- There are techniques that automate model selection

- *auto.Arima* command in R picks the best p,d & q parameters for ARIMA(p,d,q)

# Time Series Model Building Using ARIMA - RPM



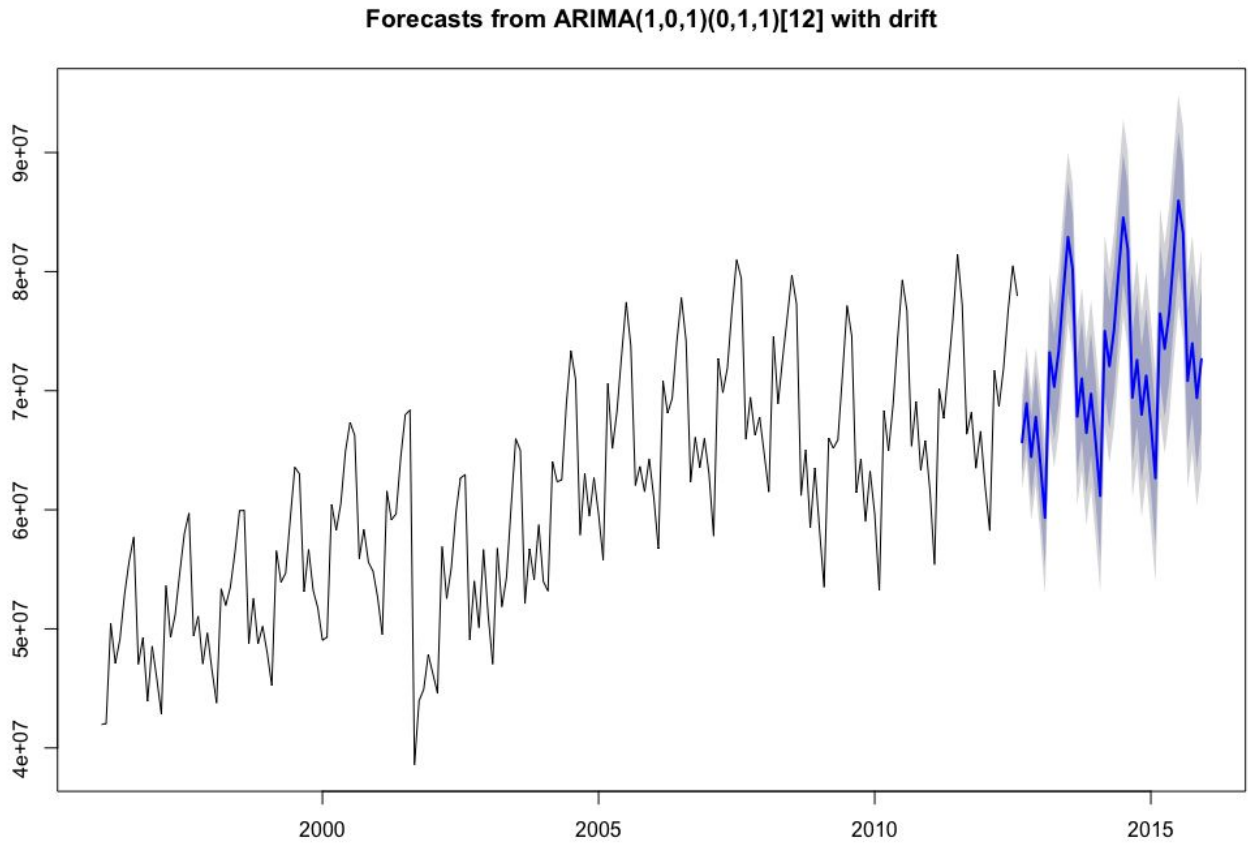
# Time Series Model Building Using ARIMA - RPM

## Auto ARIMA

```
Series: milestimeseries
ARIMA(1,0,1)(0,1,1)[12] with drift

Coefficients:
          ar1      ma1     sma1      drift
      0.9078  -0.2093  -0.7266  110280.44
s.e.  0.0364   0.0885   0.0682   31856.26

sigma^2 estimated as 3.901e+12:  log likelihood=-2994.93
AIC=5999.86   AICc=6000.19   BIC=6016.04
```

# Time Series Model Building Using ARIMA - RPM

## Forecast




MAPE: 1.65%

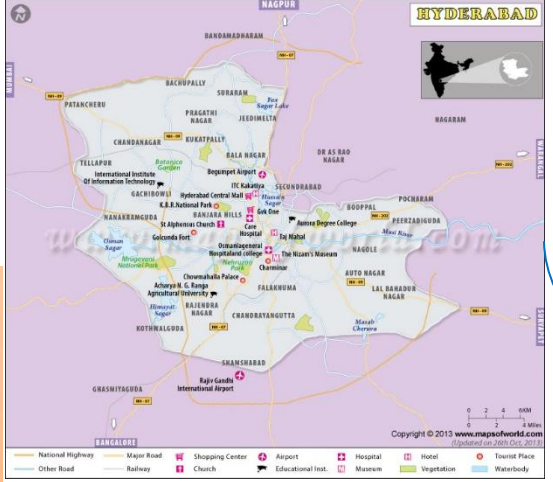
# Resources

- http://a-little-book-of-r-for-time-series.readthedocs.io/en/latest/src/timeseries.html A short condensed summary on time-series

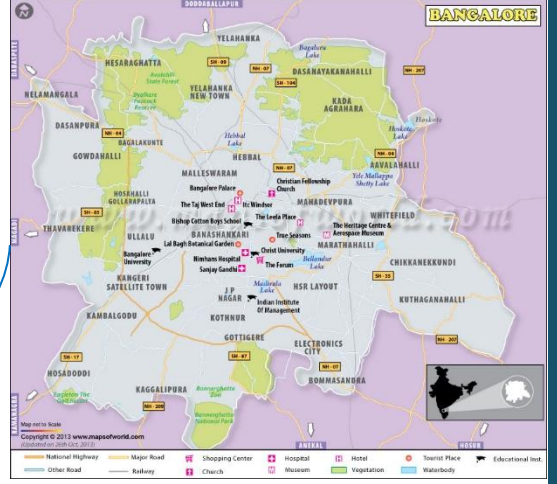
- https://onlinecourses.science.psu.edu/stat510/ Applied Time Series Analysis STAT 510 : An online course offered by Penn State Eberly College of Science

- https://www.otexts.org/fpp An good open online book on Forecasting methods and practices

- https://www.analyticsvidhya.com/blog/2015/12/complete-tutorial-time-series-modeling/ A short tutorial on using ARIMA models




Inspire...Educate...Transform.




## HYDERABAD

2nd Floor, Jyothi Imperial, Vamsiram Builders, Old Mumbai Highway, Gachibowli, Hyderabad - 500 032
+91-9701685511 (Individuals)
+91-9618483483 (Corporates)

## BENGALURU

L77, 15th Cross Road, 3rd Main Road, Sector 6, HSR Layout, Bengaluru – 560 102
+91-9502334561 (Individuals)
+91-9502799088 (Corporates)

### Social Media

| | |
|---|---|
| Web: | http://www.insofe.edu.in |
| Facebook: | https://www.facebook.com/insofe |
| Twitter: | https://twitter.com/Insofeedu |
| YouTube: | http://www.youtube.com/InsofeVideos |
| SlideShare: | http://www.slideshare.net/INSOFE |
| LinkedIn: | http://www.linkedin.com/company/international-school-of-engineering |

*This presentation may contain references to findings of various reports available in the public domain. INSOFE makes no representation as to their accuracy or that the organization subscribes to those findings.*